Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Computer and Control Engineering (38^{th}cycle)

Toward Robust, Responsible and Trustworthy Speech Foundation Models

By

Alkis Koudounas

******

Supervisor(s):
Prof. Elena Baralis, Supervisor
Prof. Eliana Pastor, Co-Supervisor

Doctoral Examination Committee:
Prof. Abolfazl Asudeh, Referee, University of Illinois Chicago, USA
Prof. Nicole Dalia Cilia, Referee, Kore University of Enna, Italy
Prof. Paolo Garza, Politecnico di Torino, Italy

Politecnico di Torino
2025

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Alkis Koudounas
2025

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

To Asya, who walked with me through every step,
and to my family, who never stopped believing in me.
To my grandma, who I know is still watching me and cheering for me.

Per aspera ad astra.

Acknowledgements

I would like to express my gratitude to my supervisor, *Elena*, for her guidance throughout this journey and for the independence she gave me to explore new directions and topics. Her trust in my ideas and her constant support have been fundamental to my growth as a researcher. I am also sincerely grateful to my co-supervisor, *Eliana*, for the innumerable hours of stimulating discussions, insightful advice, and encouragement that shaped this work and made this research possible.

I would like to thank *Moreno*, for being the best research buddy I could have hoped for: for the countless nights spent debugging, the ideas shared, the conferences experienced together, and the papers written, rewritten, turned upside down, and presented, and for the thousands of coffees we drank along the way. I am grateful to *Flavio*, for teaching me innumerable coding tricks, for constantly pushing me to refine my skills, and for showing me just how challenging mock interviews can be. I also thank *Marco*, for being my unofficial technical mentor, for always being available when I needed advice, and for helping me navigate important decisions along the way.

My thanks go to all my fellow *MINDS researchers and PhD students*, who never failed to make this path a little lighter, filling our days with laughter, discussions, and countless coffee breaks that often mattered more than we realized.

I am especially grateful to *Manuel*, who played a pivotal role in my first internship experience and allowed me to see Amazon from a unique and insightful perspective. I would also like to thank my managers and mentors at Amazon, *Simon, Deepak, Venky, Jinming*, and all the researchers I had the pleasure to work with and learn from. Their guidance, feedback, and trust have been invaluable, and these experiences have contributed enormously to both my professional and personal growth.

I would also like to thank *Matteo, Lorenzo, and Alessio*, for being wonderful friends throughout this journey. Thank you for your patience when I disappeared

for days at a time, for never holding it against me, and for celebrating every step and every success with me. Your support, understanding, and friendship have meant more to me than you know.

Finally, and most importantly, I would like to thank *my parents, sister, and new brother-in-law* for their constant and unconditional support. Regardless of my mood, anxieties, or circumstances, they have always been there for me, cheering me on and believing in me even when I struggled to do so myself. And to *my partner*, who truly stood by me every step of the way: this PhD is also, in part, yours. Thank you for your patience, understanding, and constant presence during the countless weekends and late nights I spent working, always with a smile and a gentle *"I'll go to bed and wait for you there,"* even when you knew I would stay up for a few more hours. Your quiet support, your encouragement, and your ability to make everything feel lighter carried me through moments when it mattered most.

I hope I have made you at least a little proud of me, today more than yesterday, but always less than tomorrow. Thank you for helping me grow into the person, and the researcher, I am today.

I love you all.

Abstract

Recent advances in self-supervised learning, large-scale pretraining, and end-to-end optimization have led to significant progress in speech processing. Despite these advancements, current systems face critical limitations in real-world deployment: systematic performance disparities across population subgroups, poor generalization to non-standard speech patterns, and vulnerability to errors like hallucinations. These challenges, combined with growing demands for privacy protection and responsible AI deployment, necessitate fundamental advances in how we design, evaluate, and deploy speech foundation models.

This thesis introduces comprehensive frameworks and methodologies for building robust, responsible, and trustworthy speech systems across four dimensions: (i) model fairness and privacy, through novel bias detection and mitigation techniques that maintain privacy while reducing performance disparities; (ii) systematic evaluation, through new benchmarking frameworks that enable rigorous assessment of hallucinations, privacy preservation, and cross-domain generalization; (iii) natural conversation modeling, through new foundation models and datasets that advance emotional understanding and multilingual capabilities; and (iv) clinical applications, through specialized architectures that improve pathological voice detection and reduce dysarthric speech recognition errors.

In model fairness and privacy, we develop a general-purpose divergence framework for identifying performance disparities across demographic, acoustic, and task-related subgroups. We introduce multiple mitigation strategies, including divergence-regularized fine-tuning and contrastive learning, reducing performance gaps by up to 40%. We further demonstrate effective bias mitigation without accessing sensitive attributes during deployment.

For systematic evaluation, we introduce three novel benchmarking frameworks. SHALLOW provides the first approach to quantifying and characterizing hallucina-

tions in ASR, revealing critical failure modes in high-stakes applications. UnSLU-BENCH establishes standardized protocols for evaluating machine unlearning in spoken language understanding (SLU), addressing growing privacy regulations. ARCH creates a unified platform for assessing audio representation learning across speech, music, and environmental sounds, enabling rigorous evaluation of cross-domain generalization.

In natural conversation modeling, we release voc2vec, the first foundation model for non-verbal vocalizations, achieving a 5% performance improvement across affective tasks. We complement this effort with DeepDialogue, a 1000-hour emotional multi-turn dialogue dataset with rich emotional annotations, and ITALIC, the first large-scale Italian SLU dataset, promoting and advancing both emotional understanding and linguistic diversity.

For clinical applications, we focus on two representative cases, pathological voice detection and dysarthric speech recognition. We introduce MVP, a multi-source fusion framework combining sustained vowels and continuous speech, and develop a two-stage approach to dysarthric speech recognition that reduces word error rates up to 37% on challenging datasets.

We evaluate our contributions across more than 20 datasets, multiple languages, and diverse model architectures, demonstrating consistent improvements under standard, impaired, low-resource, and privacy-sensitive conditions. To support reproducibility, we release open-source datasets, pre-trained models, and evaluation frameworks.

This thesis establishes scalable techniques for developing more equitable, reliable, and responsible speech AI systems, enabling practical deployment in critical scenarios from multilingual assistants and healthcare tools to inclusive speech recognition for diverse speaker populations. By addressing fundamental challenges in robustness, evaluation, and ethical AI development, this work provides essential building blocks for the next generation of speech technology.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context and Motivation

Speech technologies are rapidly becoming an integral part of how we interact with information, devices, and services. From the ubiquity of voice assistants in our phones and smart home ecosystems to the deployment of transcription tools in media, real-time translation services breaking down language barriers, and automated systems in customer service, speech processing now supports a vast range of applications with profound impact on society and the economy. Rapid progress in speech technology has been driven by advances in model architectures, learning methods, and data availability. The most significant change has been the adoption of self-supervised learning (SSL) and end-to-end (E2E) optimization. Models like wav2vec 2.0 [1] and HuBERT [2] demonstrate this new approach, learning directly from large amounts of raw audio without human labels. This paradigm shift has reduced the need for expensive manual data annotation and enabled better transfer learning across many tasks, ranging from basic speech recognition to more complex problems like speaker identification and emotion analysis.

Despite these advancements, the transition from controlled laboratory environments to real-world deployment remains challenging. Current state-of-the-art systems, while achieving impressive aggregate performance, often exhibit significant and systematic failures on specific groups of speakers [3, 4]. These hidden disparities can manifest across demographic factors (e.g., based on age, gender, or accent), linguistic backgrounds (e.g., non-native speakers or speakers of low-resource dialects), and

pathological subpopulations (e.g., individuals with voice disorders). Models also prove fragile when encounter non-standard inputs, struggling with the acoustic variability of dysarthric speech, the phonetic nuances of regional accents, or even simple environmental factors like background noise and microphone distortion [5, 6]. Furthermore, a particularly dangerous failure mode is hallucination, where a model generates a confident yet entirely incorrect output [7]. In speech recognition, this can lead to the insertion of phantom words or phrases, which are especially harmful because their high confidence scores make them difficult for downstream systems or human users to detect. These limitations become critical in high-stakes domains such as healthcare, legal transcription, or accessibility services, where failure is not an inconvenience but a direct path to tangible harm.

At the same time, machine learning is shifting toward making responsibility, fairness, and privacy core requirements for trustworthy AI [8–11]. In the context of speech, this paradigm shift introduces a complex set of open research questions. First, it necessitates moving beyond simple performance metrics to robust methods for bias discovery and equitable evaluation across finely-grained subgroups. Second, it demands the development of privacy-preserving model architectures, recognizing that raw audio is an incredibly rich signal containing not just spoken content but also sensitive paralinguistic information about speaker identity, emotional state, physical health, and background environment. Third, it raises the critical need for data governance, including the right for users to have their data removed from trained models post hoc, a technical challenge known as *machine unlearning* [12]. These issues become even more challenging with modern foundation models. Their large size and complex behavior make it hard to understand how they make decisions. This lack of transparency makes it difficult to ensure the models are fair, reliable, and aligned with human values.

In this thesis, we explore how to design speech foundation models that are not only performant, but also robust, responsible, and trustworthy. We argue that these attributes are not post-hoc additions but must be integrated into the core of the model design, training, and evaluation lifecycle. We advance this goal through four interconnected technical perspectives that address fundamental challenges in modern speech technology.

The first perspective focuses on equitable and privacy-preserving model behavior. We develop novel techniques for identifying and mitigating performance dispari-

ties across diverse populations through automated subgroup discovery and targeted interventions. Our approaches can also operate without requiring pre-defined demographic labels, creating learning objectives that explicitly promote fairness while protecting user privacy. This work establishes new methods for detecting and addressing systematic biases while maintaining strict privacy guarantees.

The second perspective establishes comprehensive evaluation frameworks that move beyond simple accuracy metrics. We introduce novel benchmarks for systematically assessing critical model behaviors: detecting and characterizing hallucinations in speech recognition, evaluating machine unlearning capabilities for privacy preservation, and measuring cross-domain generalization in audio processing. These frameworks enable systematic assessment of model reliability, privacy preservation, and robustness across diverse operational conditions.

The third perspective advances natural conversation modeling by expanding the scope of speech representation learning. We move beyond standard transactional speech to capture the full richness of human vocal communication, including non-verbal sounds, emotional prosody, and linguistic variations. Through new datasets and specialized foundation models, we enable more natural and inclusive speech technology that better reflects the diversity of human expression.

The fourth perspective addresses the unique challenges of clinical and impaired speech. We develop specialized architectures and fusion techniques explicitly tailored to pathological voice detection and dysarthric speech recognition. This work demonstrates how careful consideration of domain-specific requirements can dramatically improve model performance on traditionally underserved populations.

Across these four perspectives, we emphasize the critical need for evaluation beyond mere accuracy. Our guiding principle is the importance of designing systems that align with real-world constraints such as data scarcity, deployment safety, and the fundamental right to privacy. Through this comprehensive approach, we establish new methodologies for developing speech technology that serves all users reliably and equitably.

1.2 Speech Foundation Models: A (Short) Overview

Foundation models represent a major advance in speech technology, following similar progress in language and vision processing. These models start by learning from large amounts of unlabeled audio data, without needing expensive human transcription. They learn through self-supervised tasks, like predicting missing parts of audio signals or distinguishing between different sounds. This initial learning phase, called pre-training, helps the model develop a broad understanding of speech patterns. This powerful base model can then be easily adapted for many different downstream tasks, such as automatic speech recognition (ASR), spoken language understanding (SLU), speech emotion recognition (SER), or speaker identification (SID) and verification (SV).

Several well-known models demonstrate the success of this approach. Wav2vec 2.0 [1] and HuBERT [2] pioneered learning directly from the raw audio waveform. More recent models like Whisper [13] achieve strong performance on speech-to-text tasks without task-specific training. Multilingual models like XLS-R [14] can now learn from over a hundred languages simultaneously. These models work particularly well with limited data, making them valuable for languages with few resources.
However, despite their impressive success on standard tests, current foundation models often overlook key real-world challenges, leading to significant problems when they are deployed. We can identify four major limitations.

First, they learn from a narrow range of speech patterns. Most of the massive datasets used for pre-training consists of clear, fluent speech from standard speakers. This makes the models unreliable when dealing with different speech patterns. They often fail completely when processing clinical speech, such as pathological or dysarthric voices.

Second, their evaluation methods hide important problems. We typically measure a model's quality with overall metrics like word error rate (WER) or accuracy. But a single average score can be very misleading, as it hides how the model performs on different subgroups of people. A model might show good overall performance while failing badly for specific populations. This makes it hard to identify dangerous failures where models discriminate against speakers based on their age, gender, or accent.

Third, they focus too narrowly on converting speech to text. Human speech is much more than just words; it includes paralinguistic signals like tone of voice, rhythm, pauses, and vocal bursts like laughter, sighs, coughs. These aspects carry crucial information about emotion, intent, and social context. Current foundation models largely ignore these elements while focusing on transcribing the words, missing much of what makes human speech meaningful.

Fourth, they lack proper privacy protections. Voice data contains sensitive information about identity, age, health, and emotional state. Current models collect and use this information without giving users control over their data. Users cannot easily know what information models have learned about them or request its removal.

As a result of these blind spots, the models can behave in unexpected and harmful ways in real-world applications. They can amplify existing social biases, fail completely for vulnerable users who most need speech technology, and produce confident-sounding errors that are very difficult to notice or fix. These issues reduce user trust in speech technology and limit its positive impact.

This thesis addresses these challenges through four main contributions. We develop new methods to ensure models work fairly for all users, while also protecting privacy. We create better ways to evaluate model behavior beyond simple accuracy measures. We expand speech technology to handle natural conversation including non-verbal communication. We design specialized approaches for clinical applications where current models often fail. Together, these advances help create speech technology that is more reliable, fair, and trustworthy.

1.3 Research Domains and Contributions

Our work advances speech technology through four complementary research directions. We address fundamental challenges in model fairness and privacy, systematic evaluation, natural conversation modeling, and clinical applications. Each direction presents unique challenges in modeling, data collection, and practical deployment. Below, we detail our contributions in each area.

Model fairness, bias, and privacy. Speech models often perform unfairly across different speaker groups, varying by age, gender, accent, emotional tone, or speaking style. Standard evaluation metrics hide these disparities by averaging performance

across all users. A model showing good overall accuracy might still consistently fail for specific populations, creating barriers for the people who need this technology most.

We address this challenge through a comprehensive framework for discovering and mitigating performance disparities. Our approach uses statistical divergence [15] to automatically identify problematic subgroups where models underperform significantly [16–19]. Unlike traditional methods, our framework does not need pre-defined demographic categories, making it more flexible.

To mitigate these discovered disparities, we develop several complementary solutions. Our divergence-regularized fine-tuning method adjusts how models learn to ensure more equal performance across groups [20]. We introduce targeted data collection strategies that focus on gathering examples from underperforming groups [21]. We futher create new contrastive learning techniques that help models develop fairer internal representations [22].

Importantly, our methods work also without needing access to sensitive demographic information during deployment. We develop privacy-preserving techniques that can detect and address bias while protecting user privacy [23, 24]. This makes our approach practical for real-world systems where collecting demographic data might be inappropriate or restricted.

Together, these advances provide a complete solution for building fairer speech technology while respecting user privacy.

Systematic evaluation. Current evaluation methods miss critical model failures that can have serious real-world consequences. One major issue is hallucination in speech recognition, where models generate confident but completely incorrect transcriptions. These errors are particularly dangerous in applications like healthcare or legal transcription, where hallucinated content can lead to wrong decisions. We address this through SHALLOW [25], the first framework to systematically measure and categorize hallucinations across lexical, phonetic, morphological, and semantic dimensions.

Privacy preservation presents another crucial evaluation challenge. As privacy regulations strengthen, users need ways to remove their data from trained models. We introduce UnSLU-BENCH [26], the first benchmark for evaluating how effectively models can "unlearn" specific user data while maintaining their general capabilities. This framework helps develop more privacy-respecting speech technology that complies with regulations like GDPR.

Speech models also need to work reliably across many different types of audio, not just clean speech. We create ARCH [27], a comprehensive evaluation platform that tests models on speech, music, and environmental sounds. Together, these frameworks enable more thorough and meaningful evaluation of speech models, helping identify and address critical failures before deployment.

Natural conversation modeling. Human speech contains much more than just words, yet current speech technology focuses almost entirely on transcribing verbal content. Non-verbal sounds like laughter, crying, and sighs carry crucial information about emotions, intentions, and social context. Similarly, the way we speak, our tone, rhythm, and regional variations, adds layers of meaning beyond the words themselves. Current foundation models largely ignore these essential aspects of human communication.

We expand speech technology to capture this richness through several complementary contributions. Voc2vec [28] represents the first foundation model specifically designed to understand non-verbal vocalizations. By learning directly from affective sounds rather than speech, it achieves significant improvements in tasks like emotion recognition and baby cry detection. This advances our ability to capture the full spectrum of human vocal expression.

To support more natural conversational AI, we introduce DeepDialogue [29], a large-scale dataset of emotionally rich interactions. This collection of 40,150 multi-turn conversations includes carefully designed emotional progressions across 41 different domains. Each conversation comes with both text and emotionally-appropriate speech, enabling research into how emotions evolve naturally through dialogue. This resource helps bridge the gap between current task-oriented systems and more natural, emotionally aware conversation.

We also address linguistic diversity through ITALIC [30], the first large-scale spoken language understanding dataset for Italian. Beyond collecting $16,000$ utterances for intent classification, we carefully capture regional variations in Italian speech. Building on this foundation, we conduct an extensive investigation into automatically identifying Italian language varieties directly from speech signals [31]. Through innovative applications of contrastive learning and detailed analysis of regional speech patterns, we demonstrate how speech technology can preserve and respect linguistic diversity rather than enforcing standardization.

We also develop novel explainability techniques that make speech model decisions more transparent and interpretable [32]. Our framework combines word-level audio

segment attribution with paralinguistic feature analysis, providing comprehensive insights into how models process both linguistic and acoustic aspects of speech. This contribution enables better understanding of model behavior, crucial for developing trustworthy speech systems that can be reliably deployed in real-world applications. Together, these resources help create more natural and expressive speech technology that better reflects the richness of human communication.

Clinical applications. Clinical speech presents some of the most challenging conditions for speech technology. Voice disorders and speech impairments create patterns that vary greatly between individuals. Standard models, trained on typical speech, often fail completely when processing these atypical patterns. The challenge is amplified by limited data availability, as collecting and annotating clinical speech requires specialist expertise.

We develop specialized solutions focusing on two critical medical applications. First, we tackle pathological voice detection [33], where early and accurate diagnosis can significantly impact treatment outcomes. We introduce MVP [34], a novel multi-source fusion framework that analyzes both sustained vowels and continuous speech. This approach mirrors clinical practice, where doctors evaluate multiple voice tasks to make diagnoses. By combining information from different vocal tasks, our system achieves more reliable pathology detection than traditional single-source approaches. Second, we address the persistent challenge of dysarthric speech recognition. Affecting millions worldwide, dysarthria makes speech unclear and difficult to understand even for human listeners. As voice interfaces become more common, poor recognition of dysarthric speech risks excluding people who might benefit most from this technology. We develop a two-stage approach that combines acoustic modeling with generative error correction, effectively reducing word error rates by huge margins [35].

Together, these contributions help create more effective speech technology for healthcare applications.

1.4 Positioning of Our Work

Speech technology has reached an inflection point between powerful capabilities and practical limitations. While foundation models have achieved remarkable technical capabilities, their real-world impact remains limited by gaps between laboratory

performance and practical requirements. Our work addresses this challenge by re-conceptualizing how speech systems should be designed, evaluated, and deployed.

We take a fundamentally different approach from current research in several key ways. First, while most work focuses on improving accuracy for standard tasks, we take steps toward trustworthy AI by design, arguing that responsible AI principles should be considered from the earliest stages of model development. By reframing fairness and privacy as intrinsic components of system design, rather than subsequent considerations, this approach represents a significant evolution from traditional frameworks.

Second, we challenge the traditional approach of treating speech technology problems in isolation. Our work shows how fairness, privacy, evaluation, and interpretability are deeply interconnected. Rather than addressing each challenge separately, we demonstrate how advances across different domains, from conversation modeling to clinical applications, can come together to create more reliable and trustworthy speech technology.

Third, we establish new methodological standards for the field. Our evaluation frameworks move beyond simply measuring performance to *understanding* model behavior comprehensively. Rather than accepting black-box behavior, we develop techniques to make models more interpretable and controllable. This enables not just better performance, but more trustworthy and reliable systems.

The practical impact of our approach is demonstrated through several key advances:

- A comprehensive framework for automated bias identification and mitigation in speech systems, including privacy-preserving methods that operate without demographic data;
- Novel benchmarks that advance systematic evaluation through hallucination detection, machine unlearning assessment, and cross-domain generalization testing;
- Foundation models and resources that enhance natural conversation through non-verbal understanding, emotional dialogue modeling , multilingual capabilities, regional language preservation, and interpretable decision processes;

- Specialized architectures that make speech technology more accessible to underserved populations through multi-source voice pathology detection, dysarthric speech recognition, and clinically-informed fusion strategies.

Our work also opens new research directions that will help shaping the future of speech technology. We demonstrate why the field needs new evaluation approaches that align with human values and practical needs. Our contributions in privacy preservation reveals important questions about protecting user data while maintaining model performance. The challenges we address in accessibility and multilingual support highlight the need for more inclusive technology. Through our systematic evaluation frameworks, we establish foundations for developing more transparent and controllable speech systems.

This thesis provides both immediate practical solutions and longer-term strategic directions for developing more responsible speech technology. By addressing fundamental challenges in robustness, fairness, and trustworthiness, we establish new paradigms for next-generation foundation models. Our contributions support the development of systems that better serve all users while maintaining high ethical standards and practical deployment requirements.

1.5 Additional Research Activities

While this thesis focuses on speech technology, our research interests span several other areas of artificial intelligence and machine learning.

In natural language processing, we explored various challenges in legal AI, from legal entity recognition [36] to court judgment prediction [37, 38]. Our work on explaining legal models' decisions shares methodological similarities with our speech model interpretability research. We developed approaches for Italian dialect classification [39], automated literature review [40], and video content analysis [41]. In the emerging field of LLM reliability, we contributed methods for hallucination detection [42] and uncertainty quantification [43].

Our research in music technology led to advances in deep music generation through Ainur [44], a system that harmonizes lyrics with audio embeddings for high-quality music synthesis. We also explored audio analysis in other domains, including acoustic-based mosquito detection [45].

In space applications, we pursued several interconnected research directions. Our work on exoplanet atmospheric analysis [46] led to novel subgroup analysis techniques [47, 48] that parallel our approaches to speech model fairness. We analyzed astronaut communications using our divergence framework [49], demonstrating its applicability beyond standard speech contexts. We also developed deep learning methods for spacecraft pose estimation using time-of-flight cameras [50].

Additionally, we contributed to foundational machine learning research through work on logical tensor networks [51], interpretable quantum feature selection [52], and gradient-based manifold learning [53].

While these contributions lie outside this thesis's scope, they reflect our broader interest in developing robust and interpretable AI systems across different domains.

1.6 Outline of the Thesis

The remainder of this thesis is organized as follows.

Chapter 2 provides essential background on speech foundation models and their development. It reviews key architectures like wav2vec 2.0, HuBERT, and Whisper, and examines fundamental tasks in speech processing including ASR, SLU, and emotion recognition. The chapter also introduces emerging Speech Large Language Models (SpeechLLMs) and discusses their implications for the field.

Chapter 3 presents our comprehensive framework for analyzing and mitigating performance disparities in speech models. We introduce novel techniques for automated subgroup discovery and bias mitigation. The chapter also demonstrates how these methods can improve model fairness while maintaining privacy and performance.

Chapter 4 introduces three novel evaluation frameworks that move beyond simple accuracy metrics. SHALLOW provides the first systematic approach to measuring hallucinations in ASR systems. UnSLU-BENCH establishes standards for evaluating machine unlearning in speech models. ARCH creates a comprehensive platform for assessing audio representation learning across diverse domains. Together, these frameworks enable a more meaningful evaluation of speech systems.

Chapter 5 advances natural conversation modeling through several key contributions. We present voc2vec, the first foundation model for non-verbal vocalizations,

and DeepDialogue, a large-scale emotional conversation dataset. We introduce ITALIC, advancing Italian language understanding while preserving regional variations. Through detailed analysis of Italian language varieties, we further demonstrate effective identification of regional speech patterns. We also develop novel explainability techniques for speech classification models, making their decision-making processes more transparent and interpretable. Together, these contributions enable more natural, inclusive, and trustworthy speech technology.

Chapter 6 focuses on clinical applications of speech technology. We present specialized architectures for voice pathology detection and dysarthric speech recognition. Through the MVP framework and other advances, we demonstrate how careful consideration of clinical requirements can significantly improve model performance in healthcare settings.

Chapter 7 synthesizes our findings and discusses their broader implications. We examine how our contributions advance both the technical capabilities and responsible development of speech technology. The chapter concludes by identifying promising directions for future research in robust, responsible, and trustworthy speech AI.

Chapter 2

Background

This chapter introduces the technical foundations needed to understand the contributions of this thesis. We review key developments in modern speech technology, from self-supervised learning (§2.1) to foundation models (§2.2). We examine core tasks (§2.3) including automatic speech recognition (§2.3.1), spoken language understanding (§2.3.2), and speech emotion recognition (§2.3.3). We also explore how these technologies combine in emerging speech-language models (§2.4). Together, these concepts support the advances presented in later chapters. A graphic overview of the chapter is provided in Figure 2.1.

2.1 Self-Supervised Learning in Speech Processing

Early speech technology relied heavily on supervised learning, requiring human-labeled data for training. Every task needed its own carefully annotated dataset, such as word-level transcriptions for speech recognition. Creating these annotations was slow and expensive, limiting high-quality speech technology to languages with abundant resources. As a result, models worked well only for specific tasks in well-resourced languages.

Self-supervised learning (SSL) has transformed this scenario by removing the need for manual annotations [54–56]. The core innovation of SSL lies in its ability to leverage the inherent structure of speech data for learning, eliminating the need for manual annotations. They use *"pretext tasks"* that create training signals automatically from the data. One common approach masks parts of the audio signal and asks

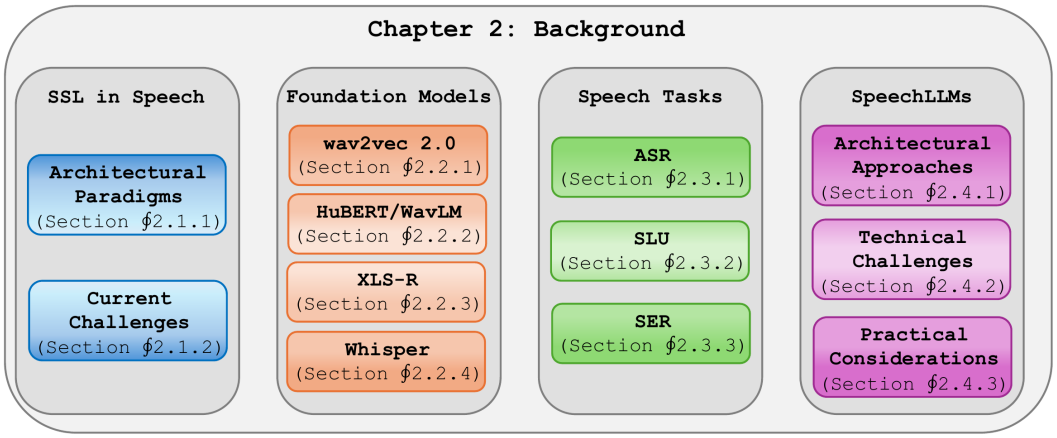


Fig. 2.1 **Chapter 2 Overview.** Graphical taxonomy of Chapter 2 topics.

the model to predict the hidden content using surrounding context [2, 57]. Through repeated exposure to thousands or even millions of hours of speech data, models develop sophisticated internal representations capturing multiple aspects of speech, including phonetics, prosody, speaker characteristics, and linguistic structure.

This paradigm shift offers two major benefits. First, models can learn from the enormous amount of unlabeled audio available online. Second, they develop versatile representations that work well across many tasks. A single pre-trained model can be adapted for different applications like speech recognition, spoken language understanding, emotion recognition, or speaker identification. This versatility makes SSL models much more practical and powerful than traditional supervised approaches.

2.1.1 Architectural Paradigms in Speech SSL

Speech processing has developed two main approaches to self-supervised learning.

The first approach uses *contrastive prediction* to learn speech patterns. It teaches models to discriminate between correct and incorrect sound sequences. The wav2vec 2.0 framework [1] demonstrates this approach through its masking and prediction system. The model first processes raw audio and masks specific parts. For each masked section, it must identify the correct sound pattern among several wrong options from the same audio. These wrong options, called "negative samples," help the model learn what makes sounds different or similar. Through repeated

comparison, the model builds a detailed understanding of speech sounds and their relationships.

The second and now dominant approach uses *masked prediction*. Inspired by the success of BERT [58] in language processing, models like HuBERT [2] and WavLM [57] predict masked audio content directly. This method works in two steps. First, it uses clustering to convert continuous audio into discrete sound units, creating automatic labels. Then, the model learns to predict these labels for masked segments using the surrounding audio context. This process helps the model understand both individual sounds and how they fit together in speech.

These models need extensive training data to work well. Early models trained mainly on LibriSpeech [59], a collection of audiobook recordings that provides clean but limited speech patterns. Newer models use more diverse data sources. They include Common Voice [60], which captures different accents and speaking styles, and real-world recordings from platforms like YouTube [13]. After this broad training, models can be fine-tuned for specific tasks like speech recognition or speaker identification.

2.1.2 Current Challenges and Research Directions

Current SSL speech models, despite their success, face several challenges that limit their real-world effectiveness. These limitations shape the research directions we pursue in this thesis.

The first challenge comes from focusing too narrowly on speech recognition performance. Both contrastive and masked prediction approaches mainly learn to identify basic speech sounds. While this works well for transcribing words, it misses other important aspects of speech. Models often ignore crucial features like emotional tone, speaking emphasis, and speaker characteristics. This narrow focus makes models less effective at understanding the full richness of human communication.

The second challenge relates to training data limitations. Current datasets, though large, lack true diversity in speaking styles and conditions. Most models train on clean, read English speech from audiobooks. This creates problems when models face real-world speech with different accents, background noise, or speaking patterns. Models particularly struggle with speech from children, elderly speakers, or people with speech impairments. The gap between training data and real-world conditions makes models unreliable in practical use.

The third challenge comes from oversimplified evaluation methods. Most evaluations rely on basic metrics like word error rate or accuracy scores. These simple measures hide important problems, such as unfair performance across different speaker groups. They miss critical failures like hallucinations, where models generate confident but incorrect transcriptions. This makes it hard to know how reliable models truly are in real-world settings.

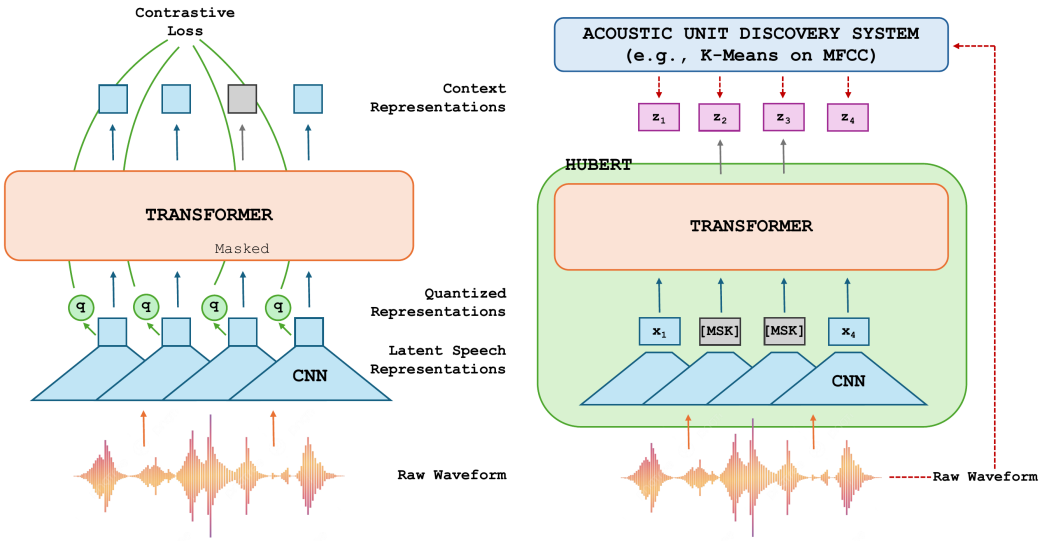
These challenges point to clear needs in speech technology development. Speech systems must effectively handle diverse speakers and speaking styles, better capture non-verbal aspects of communication, and undergo more rigorous evaluation to reveal potential problems before deployment. Addressing these fundamental challenges to create more reliable speech technology drives the research presented in this thesis.

2.2 Foundation Models for Speech Processing

Foundation models, exemplified by architectures like wav2vec 2.0 and HuBERT that we will explore in detail in Sections §2.2.1 and §2.2.2, represent a major shift in speech technology development by leveraging self-supervised learning at their core. Instead of training separate models for each task, these models first learn general speech patterns from large amounts of unlabeled data through self-supervised pre-training. After this initial training phase, they can be efficiently adapted for specific tasks through fine-tuning. This "*pre-train once, fine-tune everywhere*" approach has led to significant improvements across speech processing tasks.

Most foundation models use the Transformer architecture [61], which is particularly good at understanding patterns in sequential data. The core of these models is an encoder that learns speech representations through self-supervised training. This encoder processes raw audio using the learning techniques we discussed earlier (see §2.1.1). Through extensive training, it learns to convert sound waves into rich representations that capture many aspects of speech.

These pre-trained encoders provide a flexible foundation for many different applications. For speech recognition, we can add a decoder that converts the representations into text. For spoken language understanding, we can add classification layers that identify user intentions. This modular design makes it easy to adapt one foundation model for many different tasks.



(a) `wav2vec 2.0` architecture. (b) HuBERT architecture.

Fig. 2.2 **Model architectures I.** `wav2vec 2.0` (left) and HuBERT (right) architectures.

In the following sections, we examine several key foundation models that have shaped the field: wav2vec 2.0 (§2.2.1), which introduced contrastive learning for speech; HuBERT and WavLM (§2.2.2), which advanced masked prediction approaches; XLS-R (§2.2.3), which extended these methods to multiple languages; and Whisper (§2.2.4), which pioneered new weak supervision techniques. Each model represents a significant advance in how we process and understand speech.

2.2.1 Wav2vec 2.0

Wav2vec 2.0 [1] marked a breakthrough in self-supervised speech learning. It showed how models could learn useful speech patterns directly from unlabeled audio data. Its success established new methods for large-scale speech representation learning and influenced many later models. Most importantly, it proved that models could learn robust speech features without task-specific training.

The model uses a two-part design to process speech, as shown in Figure 2.2(a). First, a multi-layer convolutional network converts raw audio into initial representations. This feature encoder transforms sound waves into sequences of speech patterns. Second, a transformer network analyzes these patterns using self-attention mechanisms.

This combination allows the model to understand both detailed sound features and longer speech patterns.

Wav2vec 2.0 learns through contrastive training. The model looks at masked sections of audio and must identify the correct speech patterns. For each masked part, it chooses between the true pattern and several wrong options from the same audio. This approach forces the model to learn subtle differences between speech sounds. Over time, the model creates a representation space where similar sounds group together while different sounds stay separate. This training process effectively builds a comprehensive understanding of speech patterns through comparison and contrast.

The model uses a quantization module to create discrete representations of the audio signal. These quantized targets serve as stable learning objectives, helping the model develop consistent representations. The quantization process maps continuous features to a finite set of learned codebook entries, providing discrete targets for the contrastive learning task. The model then generates context representations by applying a transformer encoder to masked features, capturing broader temporal dependencies. The contrastive learning objective can be formalized as:

$$\mathscr{L}_c = -\sum_{i=1}^{T} \log \frac{\exp(sim(c_i, q_i)/\tau)}{\sum_{j=1}^{K} \exp(sim(c_i, \tilde{q}_j)/\tau)} \tag{2.1}$$

where c_i is the context representation derived from the transformer encoder, q_i is the quantized target, $\tilde{q}_j$ represents negative samples, τ is the temperature parameter, and $sim(\cdot,\cdot)$ is the cosine similarity.

Wav2vec 2.0 is trained on LibriSpeech [59], using 960 hours of English audiobook recordings. Its most impressive achievement was working well with very limited training data. The model could achieve good speech recognition results using just ten minutes of labeled speech. This breakthrough made high-quality speech recognition possible for languages with few resources. The framework's success in low-resource scenarios proved the power of the pre-training and fine-tuning approach. However, training only on clean audiobook speech leads to two major limitations. The model struggles with noisy, real-world audio and lacks robustness across diverse speaking styles.

2.2.2 HuBERT and WavLM

HuBERT (Hidden-Unit BERT) [2] represents a significant advancement in speech self-supervised learning by adapting masked language modeling to audio processing. As shown in Figure 2.2(b), it still uses a transformer architecture but introduces a new learning approach. Instead of comparing correct and incorrect patterns like wav2vec 2.0, HuBERT directly predicts the content of masked audio segments. This masked prediction approach proved more effective than contrastive methods, becoming the dominant technique in speech representation learning.

HuBERT introduces an innovative approach to self-supervised learning through its iterative refinement process. The initial phase employs k-means clustering on acoustic features to generate preliminary discrete sound units. These units serve as pseudo-labels for the first training iteration, enabling supervised-style learning without manual annotations. The trained model then generates improved representations, which facilitate more refined clustering in subsequent iterations. This iterative refinement process creates a self-improving cycle, where each stage produces increasingly sophisticated speech representations.

The masking strategy randomly masks spans of the input signal, typically covering about 40% of the audio frames. Each masked span is approximately 300ms long, chosen to roughly match the duration of phonetic units in speech. This span-based masking encourages the model to learn longer-range acoustic dependencies rather than just local patterns. The masked prediction objective is defined as:

$$\mathscr{L}_m = \sum_{t \in M} -\log p(g_t | x_{\setminus M}) \tag{2.2}$$

where M is the set of masked indices, g_t represents the target cluster assignment, and $x_{\setminus M}$ denotes the unmasked portions of the input.

HuBERT base configuration utilizes the LibriSpeech dataset as its primary training corpus, enabling direct performance comparisons with preceding models like wav2vec 2.0. However, its success inspired important improvements through WavLM [57]. WavLM enhanced the HuBERT framework in two key ways. It modified the training process and expanded the training data to include more realistic conditions. By including background noise and overlapping speech, WavLM became more robust to real-world variations.

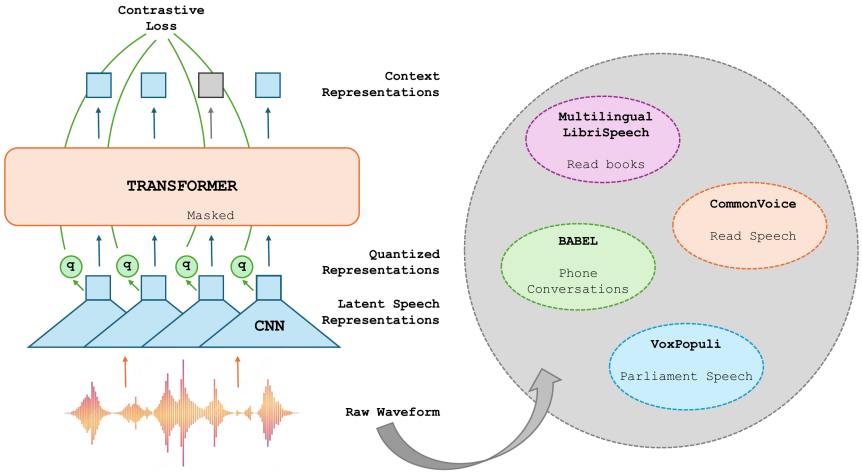


Fig. 2.3 **Model architectures II.** `XLS-R`, along with its pre-training datasets.

The primary computational consideration in implementing and pre-training these models lies in the resource-intensive nature of the multi-stage training process, particularly for larger size configurations.

2.2.3 XLS-R

XLS-R [14] exemplifies the scaling hypothesis in speech foundation models. The model investigates the effects of expanding a proven architecture, specifically wav2vec 2.0, by training it on an unprecedented scale of linguistically diverse data. The innovation lies not in architectural modifications or novel learning objectives, but in the substantial expansion of training data volume.

As shown in Figure 2.3, the model's training corpus comprises a comprehensive compilation from multiple sources, including Common Voice [60], Babel [62], and VoxPopuli [63]. The dataset encompasses 436,000 hours of speech data, equivalent to approximately 50 years of continuous audio, and spans 128 distinct languages. Despite inherent data imbalances between high-resource languages (e.g., English) and low-resource languages (e.g., Welsh, Swahili), the model learns a unified representation space across all languages.

XLS-R established itself as the predominant foundation model for multilingual and low-resource speech processing systems. It demonstrates the effectiveness of cross-lingual transfer learning. Through simultaneous learning of phonological and

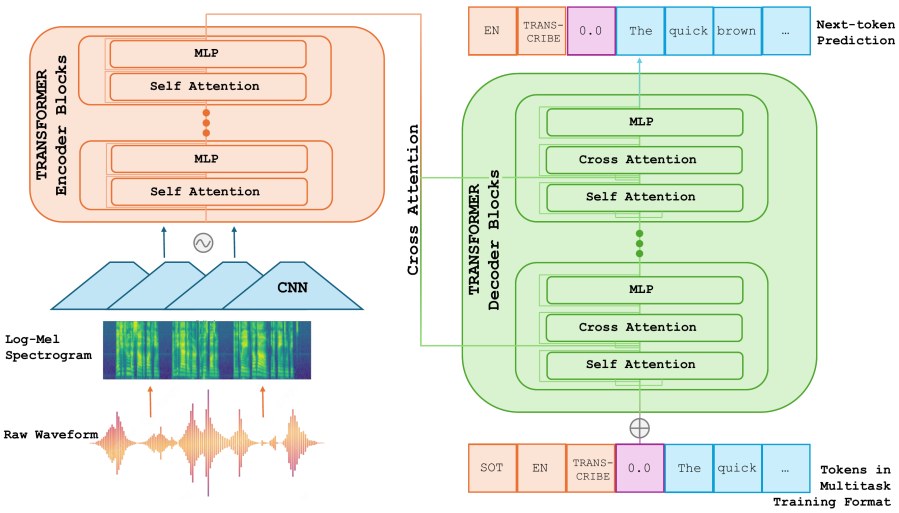


Fig. 2.4 **Model architectures III.** `Whisper`, with its encoder-decoder architecture.

structural patterns across 128 languages, the model effectively transfers knowledge from high-resource to low-resource languages during fine-tuning.

Nonetheless, while linguistically diverse, the training data predominantly consists of formal speech, lacking representation of spontaneous conversational dynamics. Furthermore, the model exhibits language interference problems due to insufficient language-specific modeling capabilities: for instance, inadvertently mixing vocabulary from multiple languages in a single transcription. Its performance often degrades on domains not covered in the training data, particularly in cross-lingual scenarios.

2.2.4 Whisper

The Whisper model [13] presents an innovative approach to speech foundation models. Unlike traditional self-supervised methods, it employs weak supervision on an extensive web-scale dataset. It distinguishes itself as an end-to-end, multitask system, departing from conventional encoders that require fine-tuning.

As displayed in Figure 2.4, the model implements a standard encoder-decoder transformer architecture. Its training corpus encompasses $680,000$ hours of internet-sourced audio data with corresponding text transcripts. The supervision is considered "*weak*" due to varying transcript quality, including potential errors, inconsistent punctuation, and audio misalignment. Despite these data imperfections, the model's exposure to diverse languages, topics, speakers, and acoustic conditions enables robust

speech understanding. The training process integrates multiple tasks simultaneously. Within a unified framework, the model performs speech-to-text transcription and translation, and language identification. This multitask training, combined with the model's large-scale pre-training (up to 1.5B parameters in its largest variant), enables strong zero-shot generalization across languages and domains. Notably, Whisper achieves competitive performance on many tasks without any task-specific fine-tuning, demonstrating the effectiveness of its approach to learning general-purpose speech representations.

The encoder-decoder architecture optimizes:

$$\mathscr{L} = -\sum_{t=1}^{T} \log p(y_t | y_{<t}, \text{enc}(x)) \tag{2.3}$$

where x is the input speech, y_t is the target token at position t, and $\text{enc}(x)$ represents the contextualized representations of the encoder.

Whisper's most notable achievement is its exceptional zero-shot capability. It demonstrates remarkable transcription accuracy across various domains and languages without task-specific fine-tuning. This immediate usability has established Whisper as a widely adopted speech recognition solution. The model exhibits several technical characteristics worth noting. It can process and infer punctuation directly from audio input. The architecture supports 99 languages officially, demonstrating strong multilingual capabilities. Its performance is particularly noteworthy in low-resource language scenarios. The model shows impressive generalization across different accents and vocabularies.

A significant limitation of Whisper lies in its tendency to hallucinate content. Studies indicate hallucination rates ranging from 1.4% to over 50% in transcriptions, particularly during longer speech pauses [64]. These hallucinations can manifest as fabricated sentences, non-existent medical procedures, or other false information. This behavior is particularly problematic in critical applications like healthcare or legal transcription. The issue appears more pronounced with non-speech audio segments and demonstrates recurring patterns in the hallucinated content [65].

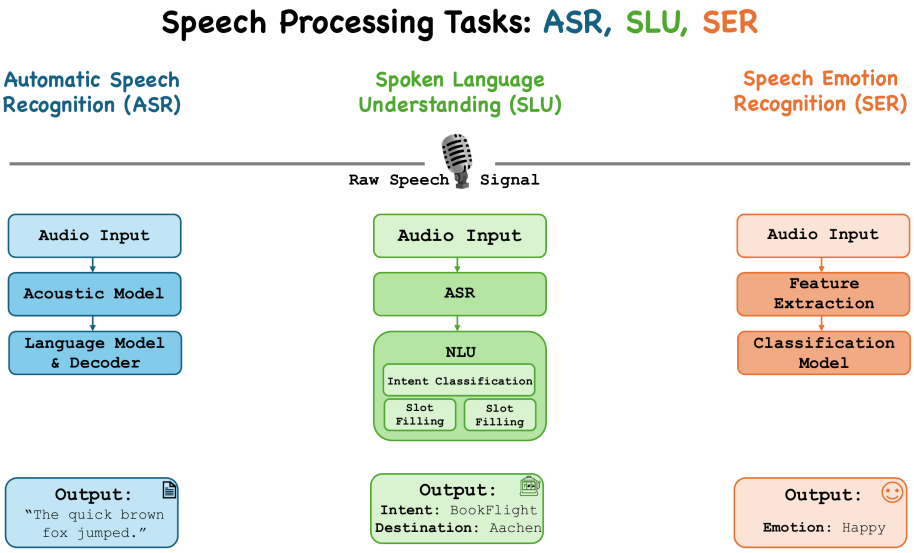


Fig. 2.5 **Standard pipeline for three speech processing tasks.** ASR (left), SLU (center), SER (right).

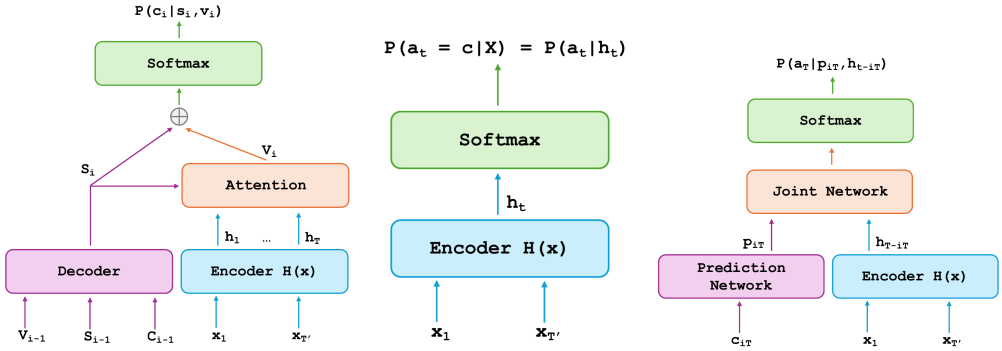
2.3 Speech Processing Tasks

Speech technology involves many different tasks that analyze, manipulate, and generate human speech signals. These tasks enable people to interact naturally with machines through voice interfaces. While each task has its specific goals, they all face common challenges: speakers talk differently, environments can be noisy, and speech signals are complex.

As shown in Figure 2.5, this section examines three core speech processing tasks: Automatic Speech Recognition (ASR) converts speech to text (§2.3.1); Spoken Language Understanding (SLU) interprets the meaning and intent behind spoken words (§2.3.2); Speech Emotion Recognition (SER) identifies emotional states from voice patterns (§2.3.3). Each task analyzes audio signals in distinct ways to capture specific information, yet collectively, they constitute the foundation of modern voice interaction systems.

2.3.1 Automatic Speech Recognition (ASR)

Automatic speech recognition converts human speech into written text. This fundamental technology bridges the gap between human communication and computer



(a) Encoder-Decoder (b) CTC-Based (c) RNN-Transducer

Fig. 2.6 **ASR Architectures.** Encoder-Decoder (left), CTC-Based (center), and RNN-Transducer (right) architectures. In the encoder-decoder model, x_i represents input audio frames, h_i are encoder hidden states, s_i is the decoder state, v_i is the context vector, and c_i is the output token prediction. The decoder takes previous predictions (c_{i-1}), states (s_{i-1}), and context vectors (v_{i-1}) as input. In the CTC-based model, x_t represents input frames, h_t are encoder hidden states, and $P(a_t = c|X)$ is the probability of outputting character c at time t given input X. In the RNN-Transducer, x_i represents input frames, h_{T-iT} are encoder states, c_{iT} is the previous output token, p_{iT} is the prediction network state, and $P(a_T|p_{iT},h_{t-iT})$ is the probability of the next output token given both acoustic and linguistic context.

processing. ASR has transformed how we interact with technology in many areas. Voice assistants and in-car systems use it for hands-free control. Professionals rely on it for tasks like medical documentation and legal transcription. It powers real-time captioning systems and voice search, making digital services more accessible.

Evolution of ASR Architectures

Speech recognition systems have undergone substantial evolution over time. Early systems used separate components working in sequence: acoustic models to identify sounds, pronunciation dictionaries to map sounds to words, and language models to ensure correct grammar. These traditional systems typically used Gaussian Mixture Models and Hidden Markov Models (GMM-HMM) for sound recognition [66–68]. While this approach had strong theoretical foundations, it has been replaced by more effective neural network solutions.

Modern ASR systems use deep neural networks that learn to convert speech to text directly [69–71]. They do not need separate components or intermediate steps.

Instead, they learn the entire process end-to-end, from raw audio to final transcription. This simpler approach achieves better results than traditional methods.

Current research focuses on three main architectures, shown in Figure 2.6. Encoder-decoder frameworks (Figure 2.6(a)) offer powerful but non-streaming recognition. The CTC-based approach (Figure 2.6(b)) provides a simple and efficient solution for speech recognition. RNN-Transducer architectures (Figure 2.6(c)) excel at real-time processing. Each design makes different trade-offs between performance, latency, and complexity.

Encoder-Decoder Frameworks. As shown in Figure 2.6(a), encoder-decoder models use a "*listen and write*" approach to convert speech to text. The encoder first processes the entire speech input, creating detailed representations that capture both sound patterns and meaning. The decoder then generates text one step at a time, using attention mechanisms to focus on relevant parts of these representations.
The attention-based decoder computes:

$$\begin{aligned} \alpha_{t,i} &= \text{Attention}(s_t, h_i) \\ c_t &= \sum_{i=1}^{T} \alpha_{t,i} h_i \\ p(y_t | y_{<t}, x) &= \text{Softmax}(W[s_t; c_t]) \end{aligned} \tag{2.4}$$

where s_t is the decoder state, h_i are encoder hidden states, $\alpha_{t,i}$ are attention weights, c_t is the context vector, and W is a learnable projection matrix.
Models like Whisper use this architecture to handle complex relationships between sounds and words, especially in long sentences. However, this approach has one main limitation: it needs the complete speech input before starting transcription. This makes it unsuitable for real-time applications where immediate responses are needed.

CTC-Based Architectures. Connectionist Temporal Classification (CTC)-based models, illustrated in Figure 2.6(b), enable streaming recognition by solving a key challenge in speech recognition: matching speech sounds to text without knowing their exact timing. These models use an encoder to analyze each small segment of audio independently. For each segment, they predict the most likely characters, including a special "blank" symbol for silence or uncertainty. The model then combines these predictions efficiently, removing repeated characters and blank symbols

to create the final text. The CTC-based model generates frame-level probabilities and optimizes the following objective:

$$\mathscr{L}_{ctc} = -\log \sum_{\pi \in B^{-1}(y)} \prod_{t=1}^{T} p(\pi_t|x) \tag{2.5}$$

where π represents all possible alignments in the set $B^{-1}(y)$ that reduce to the target transcript y after removing repeated and blank tokens, x is the input speech, and B is the mapping function that removes repeated and blank tokens (often called the CTC collapse function).
This simple approach works well with foundation models like wav2vec 2.0 and requires less computation than other methods. However, because it processes each audio segment independently, it sometimes makes mistakes in character predictions that a more context-aware approach would avoid.

Transducer Architectures. RNN-Transducer models, shown in Figure 2.6(c), are designed for real-time speech recognition where immediate response is crucial. These models use two main components working together. An audio encoder processes incoming speech frames continuously. At the same time, a prediction network tracks what text has been generated so far. A special joiner network combines information from both components to decide whether to output a new character or wait for more audio input. The transducer architecture combines acoustic and linguistic information through:

$$p(y|x) = \sum_{\hat{y} \in A(x,y)} \prod_{i=1}^{|\hat{y}|} p(\hat{y}_i|x, y_{<i}) \tag{2.6}$$

where $A(x,y)$ represents all valid alignments between input x and output y, and $\hat{y}$ is the sequence of predictions including blank symbols.
This design makes transducers ideal for live applications like captioning systems and voice interfaces. They work well on mobile devices and other systems with limited resources. The Parakeet [72] models represent the current state-of-the-art in transducer design. It uses an optimized encoder called FastConformer [73], which improves on the original Conformer [74] architecture. The Conformer is an evolution of the Transformer architecture which combines convolution operations for processing local patterns with self-attention for capturing long-range relationships in speech. FastConformer processes audio more efficiently through specialized convolution operations, making it faster while maintaining accuracy.

Performance Metrics and Evaluation Framework

Word Error Rate (WER) is the standard way to measure how well speech recognition systems perform. It compares the system's output text with human-written transcripts. WER counts three types of mistakes: substitutions (where the system picks the wrong word), deletions (where the system omits a word completely), and insertions (where the system adds words that weren't spoken). The final score is calculated as a percentage of errors relative to the total number of words in the correct transcript. In formulas:

$$\text{WER} = \frac{S + D + I}{N} \times 100\% \tag{2.7}$$

where S is the number of substitutions, D is deletions, I is insertions, and N is the total number of words in the reference transcript. Most researchers use WER because it is simple to calculate and easy to understand.

However, WER has important limitations. It only looks at whether words match exactly, ignoring their meaning and importance. For example, changing "*increase*" to "*decrease*" gets the same penalty as changing "*the*" to "*a*". This creates problems when speech recognition feeds into other systems that need to understand meaning. Recent research proposes better metrics that consider semantic similarity [75, 76], helping to identify more serious errors.

Systematic Analysis of Hallucination Phenomena

Modern ASR systems sometimes produce a dangerous type of error called hallucination. Unlike simple misrecognitions, hallucinations create fluent text that has no connection to the actual audio input [7, 11]. These errors often occur during silence, background noise, or non-speech sounds. What makes hallucinations particularly dangerous is that they sound natural and plausible, making them hard to detect.

Hallucinations can cause serious problems in critical applications. In medical settings, ASR systems might fabricate words in doctor-patient conversations or clinical notes, potentially leading to incorrect treatment decisions. In legal transcription, fabricated content could affect court decisions. In voice-controlled systems like vehicle interfaces, hallucinated commands could trigger dangerous actions. Making these errors harder to catch, systems often report high confidence in their hallucinated output.

While researchers have studied hallucinations in language models extensively, ASR hallucinations are different and less understood. We do not fully know why they happen, how different model types affect them, or how acoustic conditions trigger them. Standard metrics like WER can't properly measure these errors because they do not distinguish between simple mistakes and complete fabrications. We need better ways to analyze how acoustic features and model design contribute to hallucinations.

To address this gap, we developed SHALLOW [25], a comprehensive benchmark for analyzing ASR hallucinations. The benchmark examines multiple dimensions of hallucination, from lexical and phonetic aspects to semantic coherence. This structured approach helps identify patterns in how hallucinations form and enables systematic comparison of different ASR models. As we will detail in Chapter 4, SHALLOW provides specific metrics for measuring hallucination severity, helping develop more reliable and trustworthy speech recognition systems.

2.3.2 Spoken Language Understanding (SLU)

Spoken Language Understanding (SLU) goes beyond converting speech to text. It aims to understand what speakers actually mean and want to accomplish. This technology powers modern voice assistants and dialogue systems, helping machines understand and respond to human requests intelligently.

As shown in Figure 2.5(b), SLU systems handle two main tasks: intent classification and semantic slot filling. Intent classification figures out what the user wants to do. For example, when someone says "*Book a flight to Aachen tomorrow*," the system recognizes this as a flight booking request. Slot filling extracts specific details needed to complete the task. In the flight booking example above, it identifies *destination="Aachen"* and *date="tomorrow"* as key information. These structured details help the system take appropriate action on the user's request. The joint intent classification and slot filling objective can be expressed as:

$$p(I,S|x) = p(I|x)\prod_{t=1}^{T} p(s_t|x,y_t) \tag{2.8}$$

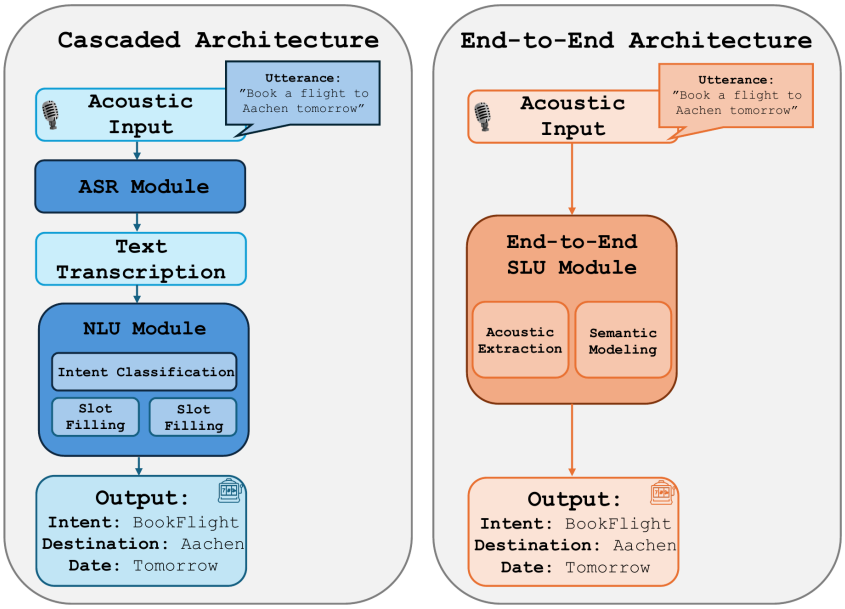


Fig. 2.7 **Spoken Language Understanding pipelines.** `Cascaded` (left), with automatic speech recognition (ASR) and natural language understanding (NLU) modules, and `end-to-end` speech-to-intent module (right).

where I is the intent, S represents slot labels, x is the input speech, y_t are the words in the transcript, s_t is the slot label for the t-th word, and T is the total number of words in the utterance.

Architectural Paradigms in SLU

As illustrated in Figure 2.7, contemporary SLU systems primarily follow two distinct architectural approaches, each presenting specific advantages and constraints.

Cascaded Architecture. The cascaded approach, shown on the left side of Figure 2.7, processes speech in two steps. First, an ASR system converts the speech to text. Then, a Natural Language Understanding (NLU) component analyzes this text to determine intent and extract relevant information. In formulas, this can be written as:

$$p(I,S|x) = \sum_{y \in Y} p(I,S|y)p(y|x) \tag{2.9}$$

where x is the speech input, y is the transcript, I is the intent, S represents slot values, and Y is the set of possible transcriptions.

This design offers important advantages. Each component can be improved independently, and the NLU module can learn from large text-only datasets. However, this sequential process has a significant weakness. Errors in speech recognition can corrupt the entire analysis. For example, if the ASR system mishears *"Turin"* as *"turning,"* the NLU component might completely misunderstand the user's intent. Additionally, important paralinguistic information like tone and emphasis is lost in the text transcription, removing potentially fundamental context for understanding user intent. These errors cascade through the system, often leaving users frustrated without clear explanation of what went wrong.

End-to-End Architecture. End-to-end systems, shown on the right side of Figure 2.7, take a different approach. They learn to map speech directly to meaning, without creating text as an intermediate step. Mathematically, we can write this as:

$$p(I,S|x) = p(I|x)\prod_{t=1}^{T} p(s_t|x,s_{<t}) \tag{2.10}$$

where x is the speech input, I is the intent, s_t represents the slot label at position t and $s_{<t}$ are the previous slot predictions.
This direct approach can capture additional meaning from how things are said, including tone and emphasis. It often handles different accents and speaking styles better because it doesn't rely on perfect transcription. By processing speech patterns directly, these systems can be more robust to variations in how people speak. However, end-to-end systems have their own limitations. Their black-box nature makes it difficult to understand or debug errors when they occur. Without intermediate text representation, developers cannot easily identify whether mistakes come from misheard speech or misunderstood meaning. These systems also typically need more training data than cascaded approaches, as they must learn both speech processing and language understanding from scratch.

Technical and Methodological Challenges

Spoken language understanding systems face several fundamental challenges that affect their real-world usefulness.

Sensitivity to input quality. These systems depend heavily on the quality of incoming speech. Different accents, background noise, and poor recording conditions can

significantly reduce accuracy. In cascaded systems, these problems become worse as ASR errors lead to further mistakes in understanding.

Low performance on low-resource languages. Many languages lack the resources needed for effective SLU systems. They face two main problems: limited transcribed speech data and few semantic annotations. This scarcity affects both acoustic modeling and semantic understanding capabilities.

Lack of fairness and bias audits. Current systems often lack systematic evaluation across different speaker groups. As a result, we do not fully understand how well they serve different populations. Even when speech recognition works well, systems may still discriminate based on demographic factors, accent or speaking style.

Our work addresses these challenges through several key contributions. We develop ITALIC, the first comprehensive SLU dataset for Italian, advancing capabilities in low-resource scenarios. Our bias analysis framework enables systematic evaluation of model behavior across different speaker groups. Through disparities mitigation techniques, we improve model fairness, while also protecting user privacy. These advances help create more inclusive and reliable SLU systems that better serve diverse user populations.

User Privacy and Machine Unlearning

SLU systems often process sensitive information in voice interactions. Users share personal details about their health, finances, and private life through voice commands [77–79]. This creates a challenging balance between improving models and protecting user privacy.

Machine unlearning has become crucial as privacy regulations like GDPR [80] and CCPA [81] become stricter, requiring the ability to remove specific user data from trained models. However, removing data from trained models presents significant technical challenges, particularly for large-scale deep learning systems where information is deeply embedded throughout the network. Traditional approaches requiring complete model retraining prove impractical for modern systems due to computational costs.

To address these challenges, we developed UnSLU-BENCH, a comprehensive evaluation framework for machine unlearning that we will present in Chapter 4. This

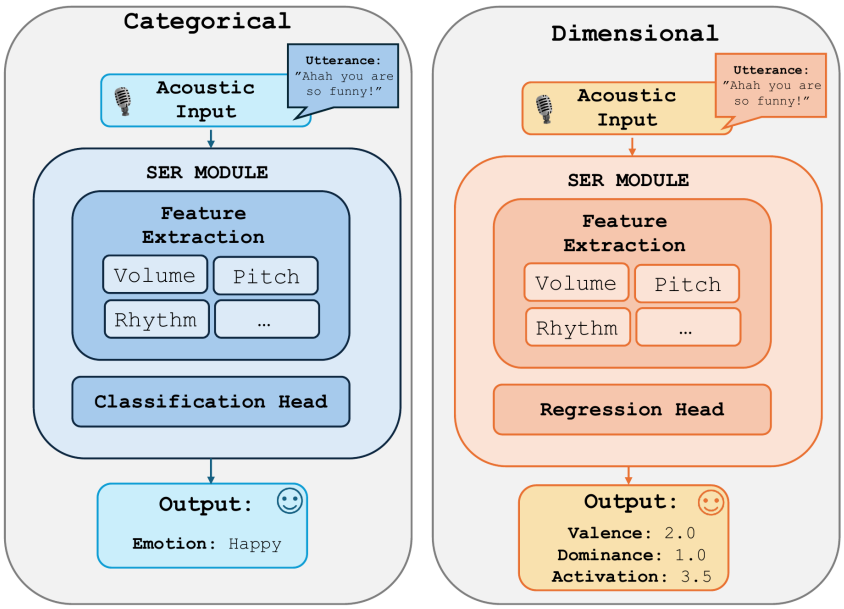


Fig. 2.8 **Speech Emotion Recognition pipelines.** `Categorical` (left) and `dimensional` frameworks.

framework helps evaluate different approaches to efficient data removal while maintaining model utility.

2.3.3 Speech Emotion Recognition (SER)

Speech contains much more than just words. It carries rich information about emotions, social interactions, and psychological states. Speakers convey these meanings through various sound patterns: changes in voice pitch, rhythm, and timing. Non-verbal sounds like laughter, sighs, and crying add another layer of emotional expression. Together, these acoustic features help listeners understand the speaker's emotions, intentions, and relationship to the conversation.

Speech Emotion Recognition (SER), illustrated in the right panel of Figure 2.5, automatically identifies emotions in speech. It analyzes multiple aspects of the voice signal, from pitch and rhythm to energy and timing patterns. This task is particularly challenging because emotions express differently across speakers and situations. Despite these challenges, SER has important applications in many areas. It helps make human-computer interaction more natural, supports mental health monitoring, and improves automated customer service. While deep learning has

advanced this field significantly, accurately capturing subtle emotional expressions remains difficult.

Theoretical Frameworks for Emotion Recognition

Researchers use two main approaches to recognize emotions in speech, each based on different psychological theories. As shown in Figure 2.8, these frameworks process speech in fundamentally different ways.

The *categorical* framework, illustrated in the left panel of Figure 2.8, treats emotions as distinct categories [82–85]. It classifies speech into basic emotional states like happiness, sadness, anger, fear, surprise, and disgust. In formulas, this approach models discrete emotion probabilities as follows:

$$p(e|x) = \text{Softmax}(f_\theta(x)) \tag{2.11}$$

where $e \in E$ is one of the predefined emotion categories and f_θ is the emotion classifier.
This approach follows the theory that humans share a set of universal, basic emotions. It is straightforward to implement and easy to understand. However, it has limitations. By forcing emotions into fixed categories, it misses subtle variations in how people express feelings. It also struggles with mixed emotions, where people feel multiple things at once.

The *dimensional* framework, shown in the right panel of Figure 2.8, takes a different approach. Instead of fixed categories, it measures emotions along three continuous scales [86–88]. *Valence* tells us if an emotion is positive or negative. *Arousal* indicates how intense or activated the emotion is. *Dominance* shows how much control or power someone feels. The dimensional approach predicts continuous values for these three dimensions simultaneously:

$$[v,a,d] = f_\theta(x), \quad \text{where } v,a,d \in [-1,1] \tag{2.12}$$

This approach can capture more subtle emotional expressions. For example, it can distinguish between different types of happiness, from calm contentment to excited joy. These three dimensions work together to describe a wide range of emotions. For instance, anger might show negative valence, high arousal, and

high dominance. Pleasure might show positive valence, low arousal, and moderate dominance. Because these dimensions are independent of each other, they can represent a wide spectrum of emotions. A single emotion can be positive but calm (like pleasure), negative but exciting (like anger), or any other combination. This flexible system helps capture the full complexity of human emotional expression. However, it requires more detailed annotation and more complex analysis methods.

Paralinguistic and Non-verbal Components

Non-verbal sounds play a crucial role in how we express emotions through speech [89, 90]. Sounds like laughter, crying, and sighs carry important emotional and social meaning without using words. However, these sounds are challenging to study and model. They are brief, often overlap with each other, and vary greatly in their acoustic patterns. Current speech models, designed mainly for understanding words, often miss these important emotional signals.

Different types of non-verbal sounds serve different communication purposes. Laughter helps create positive connections between people. Crying expresses distress or strong emotions. Sighs can show resignation, relief, or deep thought. Screams evolved as warning signals to quickly get attention in dangerous situations. Moans can mean different things depending on context, from pleasure to pain.

These sounds share remarkable patterns across different cultures. Research shows that people from different languages recognize the same basic types of vocalizations [91], advocating for their fundamental role in human communication. We even share some of these sounds with other primates, suggesting they have deep evolutionary origins [92, 93]. Each sound has specific acoustic features that consistently link to certain emotions, creating a rich system for emotional expression.

Beyond these distinct sounds, emotion also shows in how we speak. Speakers vary their tone, pitch, volume, and rhythm to add emotional meaning to their words. These features work alongside the actual words to create multiple layers of meaning. However, speakers must balance emotional expression with the rules of language. Non-verbal sounds, free from language constraints, often express emotions more directly through their acoustic patterns.

Processing these signals presents several technical challenges. Their short duration requires very precise timing analysis. When multiple sounds overlap, we need

sophisticated methods to separate them. The wide variety of possible sound patterns demands robust feature detection techniques. These difficulties help explain why traditional speech systems often ignore non-verbal sounds.

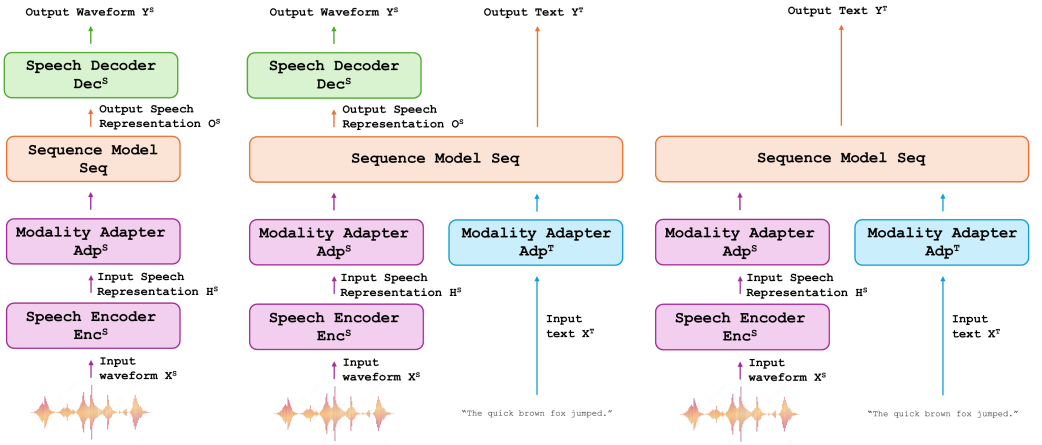
To address these fundamental challenges in processing non-verbal vocalizations, we introduce voc2vec, the first foundation model specifically designed for emotional sounds like laughter, sighs, and cries. Unlike traditional speech models that focus on words, voc2vec learns directly from emotional sounds. It learns without needing any text labels or transcriptions, focusing purely on the emotional content of sounds. This new approach helps capture the full range of human emotional expression through non-verbal communication. By understanding these important signals, voc2vec enables more natural and emotionally aware speech technology.

Emotion Recognition Datasets and Their Limitations

Several important datasets serve as benchmarks for emotion recognition research. Each offers different approaches to capturing emotional speech.

IEMOCAP [94] provides 12 hours of emotional conversations between actors. It includes both scripted and improvised interactions to capture different types of emotional expression. Ten actors participated in pairs, creating five sessions of male-female conversations. RAVDESS [95] takes a more structured approach with 24 actors, evenly split between men and women. Each person records eight different emotions: happiness, sadness, anger, fear, surprise, disgust, calm, and neutral. The dataset carefully varies how strongly these emotions are expressed. CREMA-D [96] offers a larger collection with $7,442$ recordings from 91 different actors. It covers six basic emotions and includes both voice and video recordings. The MSP-Podcast dataset [97] takes a different approach by using real podcast recordings, aiming for more natural emotions.

However, these datasets have important limitations. Acted emotions, while easier to control and record, often appear more dramatic than natural ones. Professional actors might use stereotypical expressions that do not match how people express emotions in real life. Most datasets focus on short, isolated expressions rather than longer conversations. This makes it hard to study how emotions change over time or how they flow in natural dialogue. Different datasets also use different ways to label emotions, making it difficult to compare results across studies.



(a) Pure SLMs (b) Speech-Text Joint Models (c) Speech-Aware Language Models

Fig. 2.9 **SpeechLLMs Architectures.** `Pure Speech Language Models` (left), `Speech-Text Joint Models` (center), and `Speech-Aware Language Models` (right) architectures.

To address these limitations and enable more natural emotional dialogue modeling, we present DeepDialogue, a large-scale resource for building better conversational systems. This comprehensive dataset contains more than 40,000 conversations with 1,000 hours of synthesized speech. Each conversation is carefully annotated with emotions from 20 different categories across 41 different topics. By capturing how emotions change throughout conversations, DeepDialogue helps create more natural dialogue systems. This resource enables research into how emotions flow and develop during human interactions, moving beyond the constraints of acted or isolated emotional expressions.

2.4 Speech-Language Integration through LLMs

Recent advances have led to Speech Large Language Models (SpeechLLMs), which combine speech processing with large language models. These systems go beyond converting speech to text [98]. They can understand spoken content, reason about it, and generate appropriate responses. By bridging speech and language capabilities, SpeechLLMs enable more natural and capable voice interactions.

2.4.1 Architectural Approaches

Current SpeechLLMs use different strategies to process and understand speech [99]. Figure 2.9 depicts the most straightforward implementation for each of them. Each approach offers distinct advantages and faces specific challenges. Models like SALMONN [100], Qwen2-Audio [101], and Phi-4 Multimodal [102] represent key advances in this field.

Pure Speech-Language Models. The first approach (Figure 2.9(a)) processes speech directly in the acoustic domain, without relying on intermediate text representations. Models like TWIST [103] and AudioLM [104] learn patterns directly from speech signals, similar to how language models learn patterns from text. These models typically transform the continuous speech signal into discrete units that can be processed like text tokens:

$$p(\text{speech}) = \prod_{t=1}^{T} p(s_t | s_{<t}) \tag{2.13}$$

where s_t represents the discrete speech unit at time step t obtained through acoustic tokenization, $s_{<t}$ represents all previous units in the sequence (the history), and T is the total number of discrete units in the speech sequence.
This direct approach offers several advantages. The models can capture subtle acoustic features like intonation, rhythm, and speaker characteristics that might be lost in text conversion. They learn natural speech patterns and can maintain acoustic consistency in generation tasks. However, working purely in the speech domain also presents challenges. Models often struggle with long-term semantic coherence, as maintaining meaning over long sequences proves harder without explicit linguistic structure. They also require specialized training approaches to handle the continuous nature of speech signals.

Speech-Text Joint Models. The second approach (Figure 2.9(b)) explicitly combines speech and text processing to leverage the strengths of both modalities. Models like SpeechT5 [105] and Moshi [106] recognize that speech and text offer complementary information about language. While text captures linguistic structure efficiently, speech contains additional information about prosody, emotion, and speaker identity. These models typically create a shared representation space where both speech and

text can be processed together:

$$p(\text{text}, \text{speech}) = p(\text{speech}|\text{text}) \cdot p(\text{text}) \tag{2.14}$$

This joint modeling enables the system to maintain both semantic accuracy and acoustic naturalness. The key technical challenge lies in aligning speech and text representations effectively. Models must learn mappings between acoustic patterns and their corresponding textual forms while preserving important speech characteristics. This approach proves particularly powerful for tasks requiring both modalities, such as speech translation or voice conversion. However, it typically requires paired speech-text data for training, which can be expensive to obtain at scale.

Speech-Aware Language Models. The most recent approach (Figure 2.9(c)) extends existing large language models to handle speech input directly. Models like SALMONN [100], Qwen2-Audio [101], and Phi-4 Multimodal [102] build upon the sophisticated language understanding capabilities already present in LLMs. Instead of building new language understanding capabilities for speech, these models focus on transforming speech inputs into representations that existing LLMs can process. The key insight is treating speech understanding as a representation alignment problem. These models typically process speech through specialized adaptation layers that bridge the gap between acoustic and textual representations:

$$p(\text{response}|\text{speech}, \text{context}) = \prod_{t=1}^{T} p(w_t|w_{<t}, E(\text{speech}), \text{context}) \tag{2.15}$$

where w_t is the output token at position t, $w_{<t}$ represents all previous tokens, $E(\text{speech})$ represents the adapted speech embeddings, context includes additional information such as conversation history or system prompts, and T is the length of the response sequence.

This approach offers several practical advantages. Models immediately benefit from the broad knowledge and reasoning capabilities of base LLMs. They can handle complex tasks like answering questions about speech content or generating contextually appropriate responses. Recent work like Qwen2.5-Omni [107] shows how this architecture can scale to handle multiple modalities while maintaining strong performance. However, the adaptation process presents its own challenges. Some fine-grained acoustic information may be lost when converting speech to LLM-compatible representations. The models are ultimately constrained by the text-centric nature of

their LLM backbone. Balancing efficient adaptation with preservation of important speech characteristics remains an active area of research.

2.4.2 Technical Challenges

SpeechLLMs face several critical challenges that affect their practical deployment. These challenges become particularly important in real-world applications where reliability and trustworthiness are essential.

Hallucination control. SpeechLLMs can exhibit more complex hallucination patterns than traditional ASR systems. They might generate plausible but incorrect content, combining ASR errors with LLM-style fabrications. This creates new types of errors that are particularly dangerous in critical applications. For example, a model might generate a coherent response that completely misrepresents the speech content. These hallucinations often appear more convincing because they maintain linguistic fluency while deviating from the acoustic input. Our SHALLOW benchmark specifically addresses this challenge by providing systematic ways to identify and characterize these hallucinations.

Representation alignment. Bridging the gap between speech and language representations presents significant technical challenges. Speech contains temporal, acoustic, and prosodic information that doesn't directly map to text representations. Models must learn to preserve relevant acoustic features while making the representations compatible with language model expectations. This alignment becomes particularly challenging when handling different languages, accents, or speaking styles.

Privacy and control. Speech data inherently contains sensitive information about speaker identity, emotional state, and personal characteristics. Modern SpeechLLMs train on unprecedented amounts of data, often exceeding millions of hours of speech from thousands of speakers. At this massive scale, models may unintentionally memorize personal information from their training data. Current architectures lack robust mechanisms for controlling what information is retained or how it is used. This raises important privacy concerns, especially in applications handling personal or confidential conversations.

2.4.3 Practical Considerations

The choice of SpeechLLM architecture significantly affects both performance and deployment requirements. Each approach offers different trade-offs in terms of computational resources, data requirements, and application suitability.

Computational requirements. Different architectures have varying computational needs. Pure speech-language models typically require less computation but may need specialized hardware for acoustic processing. Speech-text joint models demand more resources to process both modalities simultaneously. Speech-aware language models need substantial memory to accommodate large language models but can leverage existing optimization techniques.

Data requirements. Training requirements vary significantly across approaches. Pure speech models need large amounts of speech data but no transcriptions. Multimodal joint models require paired speech-text data, which can be scarce for many languages. Speech-aware adaptation approaches can leverage existing pre-trained language models but need carefully curated speech data for effective adaptation.

Application constraints. Different applications impose varying requirements. Real-time applications like voice assistants need models that can process speech with low latency. Healthcare or legal applications require high accuracy and reliable confidence estimates. Multilingual applications need models that can handle different languages and accents effectively.

The development of SpeechLLMs represents a significant step toward more natural and capable speech interfaces. While these models offer powerful capabilities for semantic speech processing, their deployment requires careful consideration of reliability, privacy, and control mechanisms. Continued research must focus on making these systems not just powerful, but also trustworthy and responsible.

Chapter 3

Subgroup Analysis in Speech Models: From Discovery to Mitigation

3.1 Introduction

Recent advances in self-supervised learning and large-scale pretraining have led to remarkable improvements in speech processing systems. These models achieve impressive aggregate performance metrics on standard benchmarks. However, beneath these high-level metrics often lie significant and systematic disparities in how models perform across different population subgroups. A model might achieve excellent overall accuracy while consistently failing for speakers of certain ages, genders, or accents, or in specific acoustic conditions.

These hidden biases represent more than just statistical anomalies. They constitute serious barriers to deploying truly robust and equitable speech technologies. In high-stakes domains such as healthcare, legal transcription, or emergency services, such performance disparities can lead to tangible harm for affected populations. Traditional approaches to identifying these biases have relied on analyzing predefined demographic categories or protected attributes. However, this methodology faces two critical limitations. First, it constrains analysis to a limited set of anticipated categories, potentially missing other significant sources of bias. Second, it requires collecting sensitive demographic information that may be impractical or undesirable during model deployment.

This chapter presents a comprehensive framework for *discovering*, *analyzing*, and *mitigating* performance disparities in speech models. Our approach begins with automated discovery of interpretable subgroups where models exhibit systematic failures (§3.3). Rather than relying on predefined categories, we leverage pattern mining techniques to identify combinations of features that correlate with degraded performance. These features span demographic attributes when available, but also include acoustic characteristics, speaking conditions, and task-specific metadata that can be derived directly from the speech signal. Using the concept of statistical divergence, we can precisely detect specific, human-understandable cohorts where the model's performance deviates significantly from the average. This provides a clear way to answer a critical question: "*Where exactly is my model failing, and for whom?*".

Building on this foundation of automated discovery, we introduce multiple strategies for mitigating the identified biases. The first strategies employ divergence-aware data acquisition (§3.4) and augmentation (§3.5) to strategically collect additional training data or modifying existing ones from underperforming subgroups. The third approach introduces a novel regularization term (§3.5) that adjusts the model loss function to focus on underperforming subgroups. The fourth methodology leverages contrastive learning techniques (§3.6) to directly improve model representations for challenging subgroups during training. The fifth method develops a privacy-preserving confidence modeling framework (§3.7) that can identify problematic inputs without requiring sensitive demographic information at deployment time.

We demonstrate the effectiveness of this framework through extensive experiments across multiple speech tasks, languages, and model architectures. The results show that our approach can significantly reduce performance disparities while maintaining or improving overall model performance. Importantly, we show that these improvements can be achieved even without requiring demographic data information at runtime, making the framework practical for real-world applications.

The remainder of this chapter is organized as follows. Section §3.2 reviews related work in bias detection and mitigation for speech models. Section §3.3 presents our methodology for automated subgroup discovery using statistical divergence. Section §3.4 introduces our divergence-aware data acquisition strategy for post-processing bias mitigation. Section §3.5 details our in-processing mitigation approaches, including divergence-aware regularization and targeted data augmentation. Section §3.6

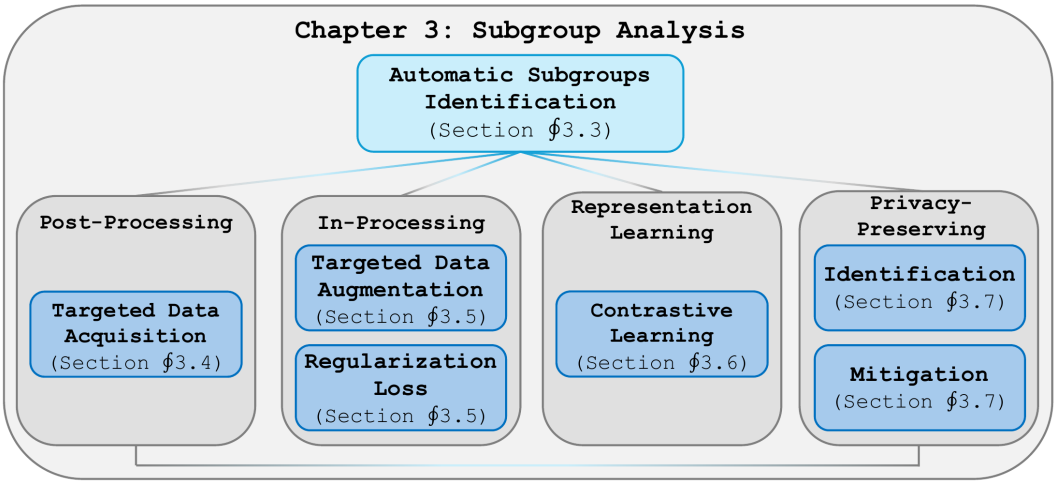


Fig. 3.1 **Chapter 3 Overview.** Graphical taxonomy of Chapter 3 topics.

presents CLUES, our contrastive learning framework for bias mitigation. Section §3.7 introduces privacy-preserving techniques for bias detection and mitigation. Section §3.8 concludes with a synthesis of our findings and their implications for developing fairer speech technology. A graphical taxonomy of this chapter's contributions is shown in Figure 3.1.

3.2 Related Work

Recent years have seen growing recognition of bias and fairness issues in speech processing systems. Multiple studies have documented systematic performance disparities across different population subgroups, particularly concerning demographic features like gender, age, accent, and ethnicity. This section reviews prior work in three key areas: automated methods for identifying problematic subgroups (§3.2.1), approaches for mitigating discovered biases (§3.2.2), and privacy-preserving techniques for deploying fairer models (§3.2.3).

3.2.1 Automated Identification of Problematic Subgroups

Early work on bias detection in speech models focused primarily on analyzing predefined demographic categories [8, 108–114]. These studies examined disparities based on individual attributes like gender, age, or accent, or specific combinations

like gender, age, and skin tone [65], gender and ethnicity [115], or intersections of demographics and geolocation [3, 116]. While valuable, this approach was limited by its reliance on anticipated categories of bias, potentially missing other significant sources of disparity.

A key advance came with the introduction of automated methods for discovering problematic subgroups. One influential approach, proposed by Dheram et al. [3], uses clustering on speaker embeddings to identify groups where models underperform. However, while automated, this clustering-based approach produces subgroups that lack direct interpretability, making it difficult to understand the root causes of performance gaps.

The automated identification of subgroups with anomalous behavior has also been explored in the context of structured data [15, 117–119]. Our approach builds on DivExplorer [15], adapting it for speech data and extending it to enable model comparison. While heuristic-driven exploration approaches like those in [118, 119] do not support model comparison, the exhaustive search of frequent subgroups offered by DivExplorer makes it uniquely suited for this task.

3.2.2 Bias Mitigation Strategies

Once problematic subgroups are identified, mitigation strategies generally fall into two categories: data-centric and model-centric approaches.

Data-centric approaches focus on improving the representation of challenging, i.e., problematic, cohorts through strategic data acquisition or augmentation. Traditional work in this area has emphasized data diversity and robustness, addressing challenges from linguistic variations and demographics [120]. Given predefined groups, researchers have explored determining optimal sample quantities using learning curves [121, 122] and achieving better coverage for underrepresented subgroups [123]. Our work differs by focusing on acquiring or modifying existing data specifically for subgroups where the model actively underperforms, regardless of their initial representation size.

Model-centric approaches modify the training process itself to address disparities. These include incorporating information from automatically identified clusters [4], and employing specialized loss functions like domain adversarial training [108, 124].

Some techniques propose addressing disparities during training through counterfactual modifications of dependent variables [111].

A particularly promising direction involves using Contrastive Learning (CL) to learn fairer representations. While CL has shown benefits for fairness in text [125, 126], image classification [127], and machine translation [128], its application to mitigating bias in speech models remains largely unexplored. Multiple works have adopted CL to train speech models [129–132], but not specifically for fairness improvement. Our work represents the first integration of contrastive loss terms to obtain improved and fairer representations for automatically identified subgroups in speech models.

3.2.3 Privacy-Preserving Methods

A significant challenge for all mitigation strategies is their reliance on sensitive demographic data, which raises privacy concerns. Recent efforts have focused particularly on protecting voice-related personal information [133–136].

Privacy-preserving techniques using speaker embeddings and group adaptation have been explored for ASR and speaker verification systems [4, 137]. However, these approaches typically rely on user-defined groups known a priori.

Our work explores a novel privacy-preserving approach built on confidence models (CMs). CMs have traditionally been used in ASR to manage errors and estimate transcription reliability [138–141]. They have also found application in downstream tasks like data selection and semi-supervised learning [141–143]. We repurpose this technology by fine-tuning a CM to identify problematic subgroups without needing sensitive metadata at inference time. This enables a privacy-preserving mitigation pipeline that can improve model fairness while respecting user privacy.

3.3 Automated Identification of Performance Disparities

This section introduces our framework for discovering and analyzing such disparities through automated subgroup identification [16, 17]. The approach enables both fine-grained performance analysis and targeted mitigation strategies while main-

taining interpretability and practical deployability. Unlike previous approaches that rely on predefined demographic categories or black-box clustering, our method identifies interpretable subgroups through pattern mining on metadata, allowing for intersectional analysis of performance gaps.

3.3.1 Subgroup Definition Through Interpretable Metadata

Speech data naturally contains or can be annotated with various types of interpretable metadata. These metadata attributes span *demographic information* (e.g., speaker gender, age), *acoustic characteristics* (e.g., speaking rate, duration of silences), *recording conditions* (e.g., noise level, environment type), and *task-specific features* (e.g., intent labels, emotion categories). Unlike approaches that rely on learned embeddings or clustering, using interpretable metadata allows us to identify and describe problematic subgroups in human-understandable terms.

We formally define a subgroup through the concept of an *itemset*, i.e., a conjunction of attribute-value pairs. Let D denote our dataset and A its set of metadata attributes. An item is an attribute equality of the form $a = v$, where $a \in A$ is an attribute and v is a possible value for that attribute. For example, if gender and age are attributes, valid items might be "*gender=female*" or "$age \in [20-40]$".

For each attribute, we require that its possible values form a partition of the dataset. This means that for attributes with continuous values, like speaking rate or duration, we must first discretize them into non-overlapping ranges. In practice, we typically use frequency-based discretization to create "*low*", "*medium*", and "*high*" categories that each contain roughly equal numbers of samples.

An itemset I is a collection of items, each referring to a distinct attribute. For example, the itemset "{*gender=female, speaking rate=high, duration=low*}" defines a subgroup consisting of short utterances by female speakers speaking at a high rate.[1] The support of an itemset, denoted $supp(I)$, is the fraction of the dataset that satisfies all its conditions. An itemset with support 0.02 represents 2% of the data. The empty itemset $\emptyset$ corresponds to the entire dataset and has support 1.

While the number of possible itemsets grows exponentially with the number of attributes, we focus our analysis on *frequent* itemsets, i.e., those whose support

[1] In what follows, we use the terms *itemset* and *subgroup* interchangeably.

exceeds a minimum threshold u. This focus serves two purposes. First, it ensures statistical significance by requiring sufficient samples to reliably measure performance Second, it maintains practical relevance by identifying disparities that affect meaningful portions of the data.

For a subset of attributes $B \subseteq A$, we denote by $I_B = \{I | attr(I) = B\}$ the set of all itemsets over attributes B. We use $I_B^{\star,u}$ to denote the subset of frequent itemsets with attributes B that meet the minimum support threshold u. When u is clear from context, we simply write $I_B^{\star}$.

3.3.2 Quantifying Performance Disparities

Given a speech model M and a performance metric f (such as accuracy or word error rate), we aim to identify subgroups where the performance of the model significantly deviates from its overall behavior. We introduce two key measures: *intra-model divergence* and *cross-model performance gap*.

Intra-Model Divergence. The intra-model divergence of an itemset I with respect to model M quantifies how much the model performance on that subgroup differs from its performance on the entire dataset:

$$\Delta_f(I,M) = f(I,M) - f(\emptyset,M) \tag{3.1}$$

where $f(I,M)$ is the average performance of model M on the subgroup defined by itemset I, and $f(\emptyset,M)$ is the model performance on the full dataset.
For example, if f measures accuracy, a negative divergence indicates worse-than-average performance on that subgroup, while a positive divergence indicates better-than-average performance. We assess the statistical significance of these divergences using Welch's t-test, which evaluates whether the difference in means between the subgroup and the overall dataset is significant. Following standard practice [144], we consider a divergence significant when the t-statistic exceeds 2.

Cross-Model Performance Gap. When comparing different models (for example, variants that differ in size, architecture, or pre-training objectives), we are interested in how their relative performance varies across subgroups. The cross-model

performance gap for an itemset I between models M_1 and M_2 is defined as:

$$gap_f(I, M_1, M_2) = f(I, M_2) - f(I, M_1) \tag{3.2}$$

This measure quantifies how much performance changes on subgroup I when replacing model M_1 with model M_2. A positive gap indicates that M_2 performs better than M_1 on that subgroup, while a negative gap indicates worse performance. As with divergence, we use Welch's t-test to assess the statistical significance of these gaps.

3.3.3 Attributing Disparities Through Shapley Analysis

Once we identify subgroups with significant divergence or performance gaps, we need to understand which specific metadata attributes contribute most to these disparities. We approach this attribution problem using concepts from cooperative game theory, specifically Shapley values [145].

Local Attribution via Shapley Values. For a given itemset I with a metric of interest $g(I)$ (either intra-model divergence or cross-model performance gap), we want to quantify how much each item $i \in I$ contributes to the total value $g(I)$. The Shapley value $s_g(i, I)$ provides this attribution by considering all possible orderings in which items could be added to form the itemset. Formally:

$$s_g(i, I) = \sum_{J \subset I \setminus \{i\}} \frac{|J|!(|I| - |J| - 1)!}{|I|!} [g(J \cup \{i\}) - g(J)] \tag{3.3}$$

where J represents any subset of I that doesn't include item i, $|J|$ is the size of subset J, $|I|$ is the total number of items in I, and $g(J \cup \{i\}) - g(J)$ measures the marginal contribution of adding item i to subset J.

The Shapley value calculation satisfies several important properties. The sum of Shapley values equals the total metric value ($\sum_{i \in I} s_g(i, I) = g(I)$), ensuring efficient attribution. Items that contribute equally to the metric receive equal attribution, maintaining symmetry. The attribution adds linearly across different metrics, enabling combination of multiple measures. Items that never change the metric receive zero attribution, properly handling non-contributing elements.

The Shapley value $s_g(i, I)$ of i in I thus measures how much a single item i contributes to the overall divergence or gap of I. A higher value means the item i *locally* contributes more to the total value $g(I)$.

Global Attribution Analysis. While local Shapley values help understand individual subgroups, we also want to assess the global impact of different metadata attributes across all subgroups. The global Shapley value $\tilde{S}_g(i)$ measures the average effect of adding item i to all compatible itemsets:

$$\tilde{S}_g(i,u) = \sum_{B \subseteq A \setminus attr(i)} \frac{|B|!(|A|-|B|-1)!}{|A|! \prod_{b \in B \cup attr(i)} m_b} \sum_{J : J \cup i \in I^{\star}_{B \cup attr(i)}} [g(J \cup i) - g(J)] \tag{3.4}$$

Here, A represents the complete set of available attributes, $attr(i)$ is the attribute of item i (e.g., "*gender*" for item "*gender=female*"), and B represents any subset of attributes not including $attr(i)$. The term m_b denotes the number of possible values for attribute b (e.g., number of possible gender categories), and $I^{\star}_{B \cup attr(i)}$ represents the set of frequent itemsets containing the attributes in B plus $attr(i)$.

The global Shapley value $\tilde{S}_g(i,u)$ provides a comprehensive measure of the influence of item i by computing its average contribution across all valid combinations of attributes. This captures how adding item i affects the performance metric g in different metadata contexts. To make this computation feasible, we consider only frequent itemsets that meet the minimum support threshold u. This global measure reveals important contextual effects. For example, while "*gender=female*"alone might show minimal impact, its effect could be substantial when combined with other attributes like speaking rate or utterance duration. For simplicity, we write $\tilde{S}_g(i)$ when the minimum support threshold u is clear from context. The complete mathematical derivation and computational details can be found in [15].

3.3.4 Efficient Subgroup Discovery

The space of possible subgroups grows exponentially with the number of metadata attributes, making exhaustive search impractical. We leverage DivExplorer [15], which implements efficient pattern mining techniques to identify frequent itemsets while computing their divergence and performance metrics.

Table 3.1 **DivExplorer subgroup exploration time.** Average (over ten runs) and worst-case execution times [s] for DivExplorer subgroup exploration. Note that FSC produces 47,736 subgroups due to its large amount of metadata.

Dataset	#Samples	#Subgroups	Avg Time	Worst Time
FSC [146]	3793	47736	1.33s	1.40s
SLURP [147]	13078	3896	0.75s	0.81s
ITALIC [30]	1441	22054	1.14s	1.19s
IEMOCAP [94]	4490	7932	1.03s	1.09s
LibriSpeech [59]	2620	2414	0.14s	0.19s

Given a minimum support threshold u (typically 0.01-0.03), DivExplorer efficiently returns all subgroups that: (i) appear frequently enough in the data (support $\geq u$); (ii) show statistically significant performance differences, and (iii) are non-redundant (not subsumed by simpler subgroups with similar divergence).

Table 3.1 shows the computational efficiency of this approach across our experimental datasets. Even for the dataset with the largest number of metadata attributes and resulting subgroups ($47,736$), the analysis completes in under 1.5 seconds.

3.3.5 Experimental Setup

Overview. Our analysis[2] spans multiple tasks, datasets, and models to ensure the generalizability of our findings. We evaluate models on Intent Classification (IC) using the Fluent Speech Commands (FSC) [146] and SLURP [147] datasets, Speech Emotion Recognition (SER) with IEMOCAP [94], and Automatic Speech Recognition (ASR) using LibriSpeech [59]. We analyze three state-of-the-art transformer-based architectures: the English-only wav2vec 2.0 [1] and HuBERT [2], and the multilingual XLS-R [14]. For wav2vec 2.0 and HuBERT, we evaluate both the base (95M parameters) and large (300M parameters) versions to examine the effect of model scale. For XLS-R, we consider only the large version, as this model is available exclusively in that configuration. For performance metrics, we use intent accuracy for IC, emotion accuracy for SER, and WER for ASR as the performance function f.

Datasets. We employ the following datasets for evaluation.

[2] github.com/koudounasalkis/Subgroup-Analysis-in-Speech-Models

Table 3.2 **Overview of the metadata used in this study.** Demographic, speech-related, and task-specific information for each dataset.

Task	Dataset	Demographics Metadata	Speech-Related Metadata	Task-Specific Metadata
IC	FSC	gender, age, country	number and duration of silences, speech rate, number of words	intent (action, object, location)
IC	SLURP	gender, country	number and duration of silences, speech rate, number of words, close/far field	intent (action, scenario)
SER	IEMOCAP (IEMO)	gender	number and duration of silences, speech rate, number of words	emotion label arousal labels (activation, valence, dominance)
ASR	LibriSpeech (LS)	gender	number and duration of silences, speech rate, number of words, number and duration of middle pauses	none

FSC. It contains 30,043 English utterances from 97 speakers, each labeled with three slots (action, object, location) that define the intent. The dataset follows a speaker-independent split across train, validation, and test sets. The test set includes 3,793 utterances from ten speakers.

SLURP. The test set comprises 13,078 utterances of task-oriented commands across diverse scenarios recorded by 142 different speakers. Each utterance is annotated with action and scenario slots, with 60 distinct intents in total.

IEMOCAP. It contains approximately 12 hours of audio-visual recordings from dyadic interactions, and divided into five sessions (i.e., splits), which are typically evaluated independently using a 5-fold cross-validation scheme. In this study, however, we consider it more appropriate to combine the test sets into a single evaluation set. This approach enables a more comprehensive assessment of model performance while increasing the size of the evaluation data. Following the standard procedure [148], we exclude imbalanced emotion categories to maintain class balance among the remaining four emotions: neutral, happy, sad, and angry. The resulting dataset comprises 4,990 samples.

LibriSpeech. We use the "*clean-360*" subset, containing 360 hours of read speech from audiobooks. The test set includes 2,620 samples from 40 distinct speakers.

Metadata selection and processing. The quality of discovered subgroups significantly depends on the choice and preparation of metadata attributes. We categorize metadata into three main types, as summarized in Table 3.2. Note that continuous attributes require discretization to create meaningful itemsets. We employ frequency-

based discretization with three bins ("*low*", "*medium*", "*high*") to ensure balanced representation while maintaining interpretability.

Demographic metadata. When available in the datasets, we incorporate speaker characteristics such as gender (available in all datasets), age (available in FSC), country of origin (available in FSC and SLURP). While these attributes are valuable for analysis during development, our framework also allows for deployment without requiring such sensitive information at inference time (see §3.7).

Speech-related metadata. We extract various acoustic features directly from the audio signal or transcripts. Specifically, we model (i) duration features, i.e., total duration of the utterance, number and total duration of silence segments, trimmed duration (excluding initial/final silences), and duration of middle pauses (for LibriSpeech[3]); (ii) speaking characteristics, i.e., speaking rate (words per second), number of words, and (for SLURP only) additional recording conditions (close/far field).

Task-specific metadata. We incorporate metadata specific to each task. For intent classification (FSC, SLURP), we include the intent slots (action, object, and location for FSC, action and scenario for SLURP). For speech emotion recognition (IEMOCAP), we include categorical emotion labels (e.g., happy, sad) and dimensional attributes (activation, valence, dominance).

Models and training. For FSC and IEMOCAP, we use publicly available fine-tuned checkpoints of wav2vec 2.0 and HuBERT from the Hugging Face hub [149]. For SLURP and LibriSpeech, we follow standard fine-tuning procedures from the literature [148]. Table 3.3 presents the performance of these fine-tuned models on each dataset. In addition, we fine-tune the multilingual XLS-R model ourselves on the FSC dataset, using both the 53- and 128-language pre-trained versions, to examine how mono- versus multilingual pre-training affects subgroup performance (Section §3.3.6d).

Table 3.3 **Overall performance of fine-tuned models.** Accuracy (%) is reported for IC and SER tasks, and WER (%) for ASR. Best results are highlighted in **bold**.

Task	Dataset	w2v2-b	w2v2-l	hub-b	hub-l
IC	FSC	91.72	93.17	98.42	**98.50**
IC	SLURP	86.86	85.59	87.69	**89.25**
SER	IEMOCAP	74.66	71.18	67.44	**74.99**
ASR	LIBRISPEECH	6.06	3.82	6.56	**3.50**

[3] Our initial analysis revealed that intermediate pause patterns (both their frequency and duration) had minimal impact on model performance across most datasets and tasks. However, these features proved particularly important for ASR performance on LibriSpeech. Given this task-specific relevance, we included pause-related metadata only for LibriSpeech, where they play a crucial role in achieving accurate transcription.

Once model fine-tuning and inference are performed, the subgroup exploration typically takes a few seconds. Table 3.1 summarizes the average and worst-case time for subgroup exploration with our approach on each dataset.

For all experiments, we set the user-defined minimum support threshold u equal to 0.03 to ensure that all subgroups in our datasets are well-represented. For the smallest test set, LibriSpeech, the smallest subgroups will include at least 75 instances. This cardinality aligns with the standard practice requiring between 50 to 100 instances for reliable results [150].

3.3.6 Results and Discussion

Our experimental analysis aims to understand and quantify performance disparities in speech models across different subgroups of data. We focus on two types of performance gaps as introduced in §3.3.2: *intra-model performance gaps*, i.e., significant differences between the performance of a model on specific subgroups versus its overall performance, quantified using the divergence metric; and *cross-model performance gaps*, i.e., differences in how various models (varying in size, architecture, or pre-training objectives) perform on the same subgroups.

Through our analysis, we address four key research questions:

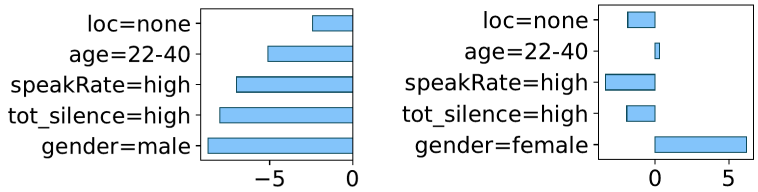
- *RQ1*: How can we automatically identify and characterize problematic subgroups for specific combinations of model, dataset, and task?
- *RQ2*: How does model size affect subgroup performance, and does the common assumption that "*larger is better*" hold true?
- *RQ3*: To what extent do performance disparities depend on model architecture?
- *RQ4*: Do multilingual models show different patterns of subgroup disparities compared to monolingual ones?

RQ1: Model Performance Across Subgroups

We begin by addressing RQ1, examining how to automatically identify subgroups where models exhibit significant performance disparities. This analysis is funda-

Table 3.4 **RQ1.** Intra-model performance gap for measure f, accuracy for FSC, SLURP, IEMOCAP (IEMO), and WER for LibriSpeech (LS), between the most negatively (S^-) and positively (S^+) divergent subgroups relative to overall test performance, using the wav2vec 2.0 base model. The *t* column reports the Welch's t-test statistic.

Dataset		Subgroups	Sup_{Train}	Sup_{Test}	f	Δ_f	t
FSC	S^-:	{*"age=22-40, gender=male, location=none, speaking rate=high, tot silence=high"*}	0.03	0.04	60.50	-31.22	7.05
	S^+:	{*"age=22-40, location=washroom, speaking rate=low, trimmed duration=high"*}	0.03	0.03	100.0	8.28	9.74
SLURP	S^-:	{*"action=quirky"*}	0.04	0.05	67.37	-19.50	10.27
	S^+:	{*"gender=female, scenario=weather"*}	0.03	0.03	95.93	9.07	8.32
IEMO	S^-:	{*"label=happy, activation=low"*}	0.03	0.03	44.74	-29.92	7.37
	S^+:	{*"label=sad, valence=low, tot silence=low, trimmed duration=high"*}	0.03	0.03	98.57	23.92	17.01
LS	S^-:	{*"gender=female, trimmed speaking rate=high, trimmed duration=low, num pauses=low"*}	0.05	0.03	17.30	11.24	4.16
	S^+:	{*"gender=female, speaking rate=low, num pauses=low, trimmed speaking rate=low, tot duration=medium"*}	0.03	0.03	3.27	-2.79	5.57



(a) w2v2-b. $\Delta_{acc} = -31.22\%$ (b) w2v2-b. $\Delta_{acc} = -0.65\%$

Fig. 3.2 **RQ1, FSC.** Contribution of individual items to accuracy, measured via Shapley values, for (a) the subgroup with the largest negative divergence (*Sup* = 0.03) and (b) the effect of considering female gender (*Sup* = 0.04) instead of male.

mental for understanding where models may fail systematically, even when their aggregate performance appears strong.

For each dataset, we analyze the wav2vec 2.0 base model's performance across all possible subgroups meeting our minimum support threshold (0.03). We consider a subgroup problematic when its performance divergence is both large in magnitude and statistically significant ($t > 2$). Table 3.4 presents the most significant positive and negative divergences found for each dataset.

Intent Classification (FSC). For the FSC dataset, we examine divergence accuracy across subgroups, where higher values indicate better performance and negative divergence indicates worse-than-average performance (first block of Table 3.4). The wav2vec 2.0 base model shows its poorest performance for the subgroup {"*age=22-*

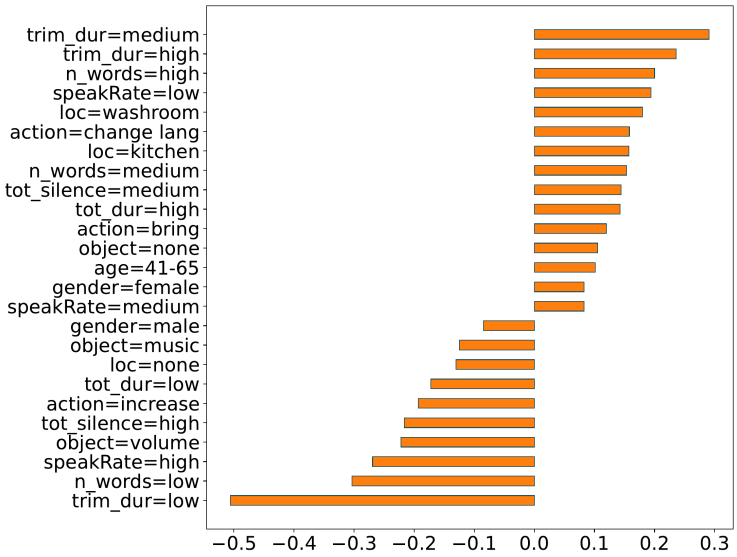


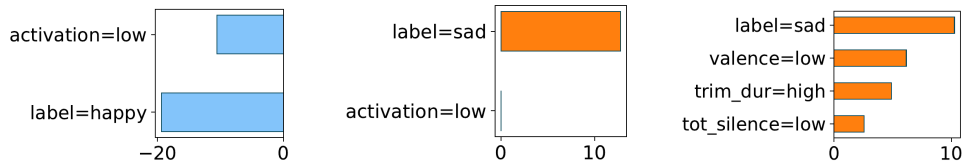
Fig. 3.3 **RQ1, FSC.** Global Shapley values for accuracy divergence, FSC, wav2vec 2.0 base. Negative Shapley values indicate items that contribute to below-average accuracy.

40, gender=male, location=none, speaking rate=high, tot silence=high"}, with a divergence $\Delta_{acc} = -31.22\%$.

Sensitive attributes like gender have a marked effect. When we consider the same subgroup but replace male gender with female while keeping other attributes constant, the performance significantly improves. This gender-based performance disparity is confirmed by the Shapley values shown in Figure 3.2, where the male gender contributes negatively to accuracy while the female gender shows a positive impact.

Conversely, we also identify subgroups where the model excels. The most positively divergent subgroup consists of utterances from speakers aged 22-40 with low speaking rate and long duration, targeting the "*washroom*" location. For this subgroup, the model achieves perfect accuracy.

The global Shapley values shown in Figure 3.3 highlight broader patterns in how different attributes affect model performance. Speaking conditions stand out as particularly influential: shorter utterances, fewer words, and faster speaking rates tend to reduce performance, whereas longer utterances, more words, and slower speaking rates improve it. These trends are consistent with known factors that affect error rates [151]. Certain intent targets also have consistent effects: references to "*volume*" are associated with lower accuracy, while mentions of the "*washroom*" location correlate with higher accuracy.



(a) Most negatively divergent subgroup. acc=44.74%, sup=0.03.

(b) Same subgroup as (a), with "sad" label. acc=87.41%, sup=0.15.

(c) Most positively divergent subgroup. acc=98.57%, sup=0.03.

Fig. 3.4 **RQ1, IEMOCAP.** Contribution of individual items to performance, measured using Shapley values, for wav2vec 2.0 base.

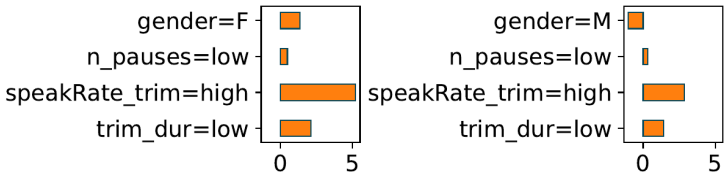
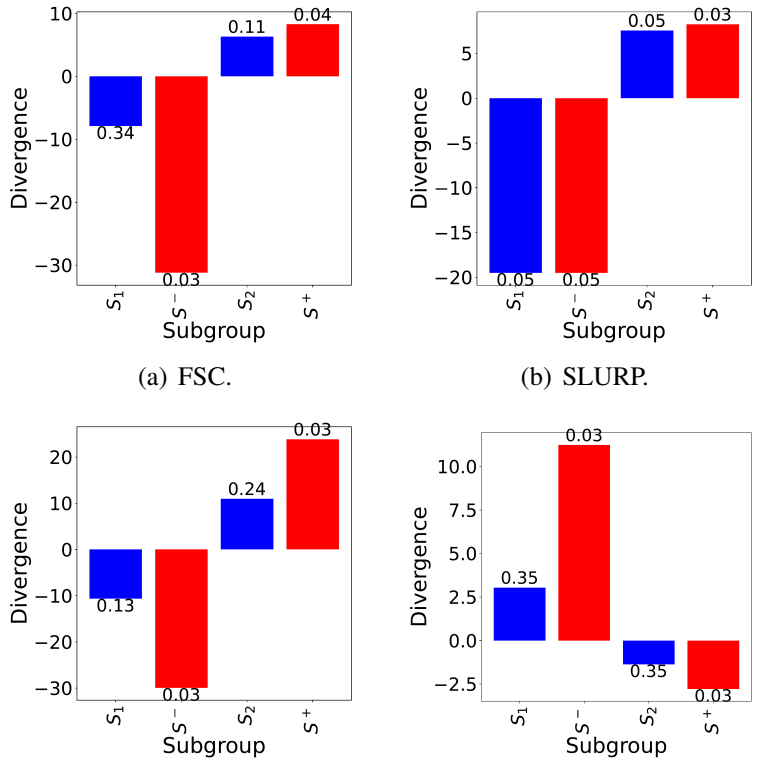


Fig. 3.5 **RQ1, LibriSpeech.** Item contributions to performance within the same subgroup, comparing *gender=female* (left; WER = 17.30%, support = 0.03) and *gender=male* (right; WER = 9.89%, support = 0.04). Wav2vec 2.0 base.

Intent Classification (SLURP). The analysis of the SLURP dataset reveals that even single attributes can drive significant performance disparities (second block of Table 3.4). The subgroup {*"action=quirky"*} experiences the highest accuracy drop, with $\Delta_{acc} = -19.50\%$. This finding demonstrates how certain intent categories alone can challenge the model, regardless of other factors.

Emotion Recognition (IEMOCAP). The IEMOCAP analysis (third block of Table 3.4) reveals complex interactions between emotion labels and other attributes. The "*happy*" emotion label, particularly when combined with low activation, associates with significantly degraded performance. The subgroup {"*label=happy, activation=low*"} shows indeed the lowest performance, with the Shapley values in Figure 3.4(a) confirming the strong negative impact of both attributes.

Interestingly, changing the emotion label to "*sad*" while maintaining other attributes (Figure 3.4(b)) leads to a huge performance improvement (accuracy rising to 87.41%). This suggests the model has particular difficulty with specific emotion-activation combinations rather than with activation levels alone. Figure 3.4(c) further demonstrates the predominant positive role of the "*sad*" label compared to other attributes in the most positively divergent subgroup.



(a) FSC. (b) SLURP.

(c) IEMOCAP. (d) LibriSpeech.

Fig. 3.6 **RQ1, Baseline Comparison.** Comparison with a baseline using the most divergent one-level subgroups. Blue bars represent baseline performance (S_1 and S_2), while red bars correspond to our approach (S^- and S^+). Wav2vec 2.0 base. Numbers above and below the bars indicate the support of each subgroup.

Automatic Speech Recognition (LibriSpeech). For LibriSpeech (last block in Table 3.4), we analyze WER, where positive divergence indicates worse performance (unlike accuracy in previous tasks). The most problematic subgroup is {"*gender=female, trimmed speaking rate=high, trimmed duration=low, num pauses=low*"}, showing significantly higher error rates.

Gender again emerges as a crucial factor. When keeping all other conditions constant but changing gender to male, the WER improves substantially (from 17.30% to 9.89%). The Shapley values (Figure 3.5) confirm the positive influence of male gender, associated with lower WER. This aligns with existing literature on ASR and speaker recognition [152–154], where male speakers and those with lower speaking rates typically see better performance.

Comparison with traditional single-attribute analysis. While conventional approaches to subgroup analysis typically focus on performance disparities across individual attributes like gender or speaking rate in isolation, our methodology examines intersectional effects by considering multiple attributes simultaneously. This more comprehensive approach reveals nuanced patterns of model behavior that single-attribute analysis might miss entirely.

To quantify this advantage, we compare the divergence scores identified by our multi-attribute method against those found using traditional single-attribute analysis. Figure 3.6 presents this comparison, demonstrating that our approach consistently identifies larger and more significant performance disparities. Across FSC, IEMOCAP, and LibriSpeech, the observed differences are pronounced, with our method detecting divergences far greater than those captured by single-attribute analyses. While SLURP shows more comparable results between the two approaches, this is only because its most problematic subgroups happen to be characterized by single attributes.

The superiority of our intersectional approach holds across all model variants we tested, including wav2vec 2.0 large and both base and large versions of HuBERT (detailed results available in the original paper [17]). These findings emphasize that understanding model biases requires looking beyond simple demographic categories to examine how different attributes interact to affect model performance. By considering these interactions, we can better identify and address performance disparities that might otherwise go unnoticed.

Summary of findings. We identified subgroups that show large deviations from the average model performance. We also determined the main sources of errors for these subgroups. By analyzing the impact of metadata values on divergence at both local and global levels, we obtained useful insights for model debugging and understanding. Our study revealed that gender has a notable effect on performance, demonstrating the value of our approach for fairness evaluation. We showed that considering multiple metadata values together, rather than individually, helps detect highly divergent behaviors. These results indicate that the proposed approach is effective in identifying subgroups with significant performance gaps. They also highlight the potential of this method to improve the performance and reliability of the models analyzed.

Table 3.5 **RQ2.** Performance gap for measure f, WER for LIBRISPEECH and accuracy for the other datasets, when scaling wav2vec 2.0 from base (90M parameters) to large (300M parameters). Arrows indicate changes: (↑) for the largest improvement, (↓) for the largest decrease. The t column shows the Welch's t-test statistic.

Dataset		Subgroups	Sup	gap_f	$f_{\texttt{w2v2-b}}$	$f_{\texttt{w2v2-l}}$	t
FSC	↑	{*"action=increase, location=none, tot duration=low, trimmed speaking rate=low, trimmed duration=low"*}	0.03	22.69	75.63	98.32	5.37
	↓	{*"action=activate, gender=male, speaking rate=low"*}	0.03	-20.97	96.77	75.81	4.92
SLURP	↑	{*"gender=female, speaking rate=high, trimmed speaking rate=high, trimmed duration=low"*}	0.04	4.08	83.88	87.96	1.83
	↓	{*"action=remove, num words=low"*}	0.03	-9.74	92.64	82.90	4.33
IEMO	↑	{*"label=happy, trimmed speaking rate=low"*}	0.04	12.96	67.28	80.25	2.66
	↓	{*"label=sad, trimmed speaking rate=low"*}	0.03	-19.86	70.55	50.68	3.53
LS	↑	{*"gender=female, num pauses=low, trimmed speaking rate=high, trimmed duration=low"*}	0.03	-5.97	17.30	11.33	1.78
	↓	{*"gender=male, num pauses=low, tot duration=low, trimmed speaking rate=high, trimmed duration=low"*}	0.04	0.46	10.17	10.64	0.14

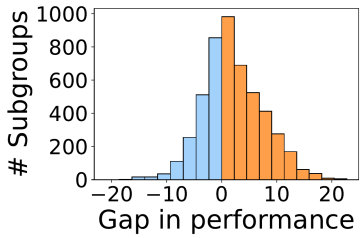
RQ2: Model Scale Impact

To address RQ2, we examine how model size affects performance disparities by comparing wav2vec 2.0 base and large variants across all datasets. Table 3.5 summarizes these comparisons, highlighting subgroups with the largest performance improvements (↑) and decreases (↓).

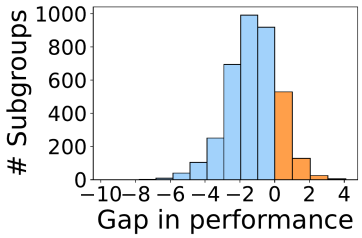
IC. For FSC, increasing model size improves overall accuracy (from 91.72% to 93.17%, Table 3.3). However, this improvement is not uniform: while 63.75% of subgroups show better performance, 31.89% actually perform worse with the larger model (Figure 3.7(a)). The highest improvement (22.69%) occurs for the subgroup {"*action=increase, location=none, tot duration=low, trimmed speaking rate=low, trimmed duration=low*"}. Conversely, the subgroup {"*action=activate, gender=male, speaking rate=low*"} shows the largest performance drop (-20.97%).

SLURP presents a different pattern. Despite similar overall performance between base and large models (86.86% vs 85.59%), the base model outperforms the large variant on 80.11% of subgroups. Figure 3.7(b) shows this distribution, with a peak around -1% performance change.

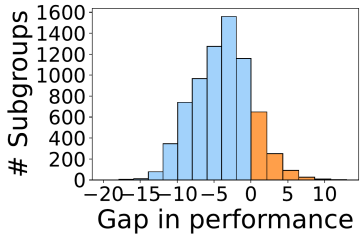
SER. For IEMOCAP, scaling up model size has a clearly detrimental effect both overall and at the subgroup level. The accuracy drops from 74.66% (base) to 71.18%



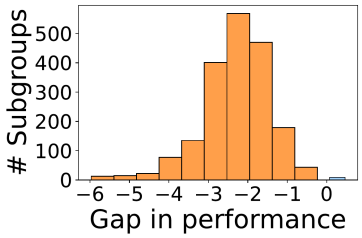
(a) FSC. Performance increase for **63.75%** of subgroups, decrease for **31.89%** of them.



(b) SLURP. Performance increase for **16.86%** of subgroups, decrease for **80.11%** of them.



(c) IEMOCAP. Performance increase for **11.21%** of subgroups, decrease for **85.85%** of them.



(d) LibriSpeech. Performance increase for **99.25%** of subgroups, decrease for **0.75%** of them.

Fig. 3.7 **RQ2.** Contribution to performance gaps when scaling wav2vec 2.0 across different datasets.

(large), with 85.85% of subgroups showing decreased performance. The subgroup {"*label=sad, trimmed speaking rate=low*"} exhibits the largest performance drop (−19.86%), while {"*label=happy, trimmed speaking rate=low*"} shows the only notable improvement (12.96%). The gap distribution (Figure 3.7(c)) shows a clear negative skew, peaking around −3%.

ASR. LibriSpeech presents the most uniform benefit from increased model size. The WER improves from 6.06% (base) to 3.82% (large), with 99.25% of subgroups showing better performance. The largest improvement (−5.97%) occurs for the subgroup {"*gender=female, num pauses=low, trimmed speaking rate=high, trimmed duration=low"*}, though this improvement is not statistically significant. Figure 3.7(d) shows the gap distribution peaking around −2.5%, indicating consistent improvements across subgroups.

Summary of findings. For the LibriSpeech dataset, we show that increasing the model size improves both overall and subgroup performance. In contrast, for the IEMOCAP dataset, a larger model results in lower performance overall and within subgroups. Our evaluation of the FSC dataset indicates that the effect of a larger

Table 3.6 **RQ3.** Performance gap for measure f when changing model architecture from wav2vec 2.0 to HuBERT base. Arrows indicate changes: (↑) largest improvement, (↓) largest decrease. The t column reports the Welch's t-test statistic.

Dataset		Subgroups	Sup	gap_f	$f_{\texttt{w2v2-b}}$	$f_{\texttt{hub-b}}$	t
FSC	↑	{*"gender=male, location=none, num words=low, tot silence=high, trimmed duration=low"*}	0.03	31.20	64.00	95.20	6.53
	↓	{*"action=decrease, age=22-40, location=washroom"*}	0.03	-1.68	100.00	98.32	1.01
SLURP	↑	{*"field=far, gender=male, tot duration=high, tot silence=low, trimmed duration=high"*}	0.03	5.46	80.76	86.22	2.13
	↓	{*"field=far, gender=female, speaking rate=low, tot duration=low, tot silence=low"*}	0.04	-3.27	85.81	82.53	1.35
IEMO	↑	{*"activation=high, label=anger, duration=low, valence=low"*}	0.03	7.54	75.34	82.88	1.57
	↓	{*"label=sad, trimmed speaking rate=low"*}	0.03	-30.14	70.55	40.41	5.41
LS	↑	{*"num pauses=medium, speaking rate=medium, tot duration=medium, tot silence=medium"*}	0.04	-1.05	7.44	6.39	0.75
	↓	{*"gender=male, num pauses=low, num words=low, tot silence=medium"*}	0.03	2.50	7.60	10.11	1.33

model can vary depending on the specific subgroup. Additionally, results on the SLURP dataset show that achieving similar overall performance does not guarantee similar performance at the subgroup level.

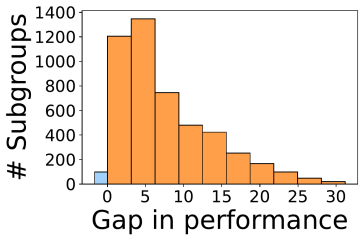
RQ3: Impact of Model Architecture

To address RQ3, we compare wav2vec 2.0 base with HuBERT base across all datasets. Table 3.6 summarizes these architectural comparisons.

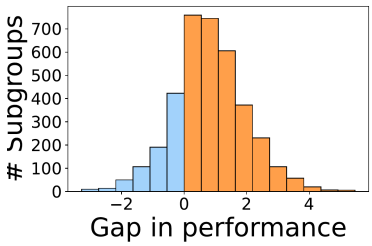
IC. For FSC, changing to HuBERT architecture yields nearly universal benefits. HuBERT base significantly outperforms wav2vec 2.0 base (98.42% vs. 91.72%), with 97.03% of subgroups showing improvement. Figure 3.8(a) shows this distribution, with a clear peak around +5% improvement.

SLURP shows more mixed results. While HuBERT performs better overall, the improvements are less uniform: 77.16% of subgroups improve, but 16.86% show degraded performance (Figure 3.8(b)). The largest improvement (5.46%) occurs for the subgroup {"*field=far, gender=male, tot duration=high, tot silence=low, trimmed duration=high*"} (t=2.13).

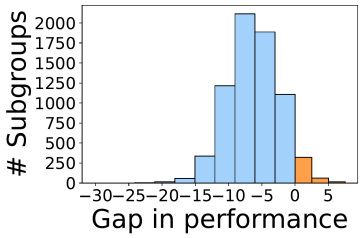
SER. IEMOCAP shows a strong preference for wav2vec 2.0 architecture. HuBERT base performs substantially worse than wav2vec 2.0 base (67.44% vs. 74.66%), with



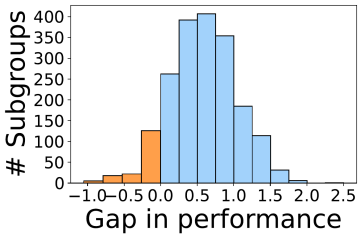
(a) FSC. Performance increase for **97.03%** of subgroups, decrease for **1.04%** of them.



(b) SLURP. Performance increase for **77.16%** of subgroups, decrease for **18.99%** of them.



(c) IEMOCAP. Performance increase for **4.15%** of subgroups, decrease for **93.95%** of them.



(d) LibriSpeech. Performance increase for **10.23%** of subgroups, decrease for **89.77%** of them.

Fig. 3.8 **RQ3.** Distribution of performance gaps when switching from wav2vec 2.0 base to HuBERT base.

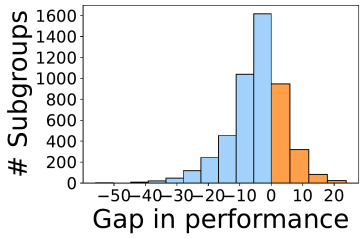
93.95% of subgroups showing decreased performance (Figure 3.8(c)). The most severe degradation (-30.14%) occurs for {"*label=sad, trimmed speaking rate=low*"} (t=5.41).

ASR. For LibriSpeech, the architectural change shows minimal overall impact (WER: 6.06% wav2vec vs. 6.56% HuBERT). However, 89.77% of subgroups show some performance decline (Figure 3.8(d)), though these changes often lack statistical significance.

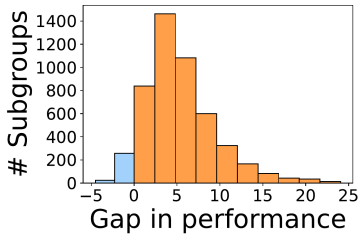
Summary of findings. We find that changing the architecture from wav2vec base to HuBERT base improves performance for almost all subgroups in FSC and for most subgroups in SLURP. In contrast, the same change reduces performance for most subgroups in IEMOCAP and LibriSpeech. These results highlight the varied impact of model architecture and the complex relationship between architecture and performance. Our findings demonstrate the limitations of comparing models only at the overall level and emphasize the importance of evaluating subgroup performance across different architectures.

Table 3.7 **RQ4.** Performance gap resulting from switching the pre-training objective from mono- to multi-lingual (XLS-R 53 and XLS-R 128, respectively), on the FSC dataset. (↑) and (↓) denote the largest increase and decrease, respectively. The *t* column shows the Welch's t-test value.

Dataset		Subgroups	Sup	gap_f	f_{mono}	$f_{\text{XLS-R 53}}$	t
FSC	↑	{*"action=activate, gender=male, trimmed speaking rate=low"*}	0.04	23.84	74.83	98.68	6.39
	↓	{*"action=increase, trimmed speaking rate=medium, object=heat, trimmed duration=high"*}	0.03	-55.83	96.67	40.83	11.49
		Subgroups	**Sup**	gap_f	f_{mono}	$f_{\text{XLS-R 128}}$	**t**
	↑	{*"gender=male, tot silence=high, location=none, trimmed duration=low, speaking rate=high"*}	0.03	24.06	75.19	99.25	6.14
	↓	{*"action=increase, object=heat, age=22-40, gender=male, tot silence=low"*}	0.04	-4.51	98.50	93.98	1.79



(a) Mono- to XLS-R 53. Performance increase for **25.59%** of subgroups, decrease for **72.30%** of them.



(b) Mono- to XLS-R 128. Performance increase for **92.90%** of subgroups, decrease for **4.05%** of them.

Fig. 3.9 **RQ4, FSC.** Gap distribution when changing the pre-training objective from mono- to XLS-R 53 (a) and from mono- to XLS-R 128 (b).

RQ4: Multilingual vs. Monolingual Models

To address RQ4, we compare three variants on the FSC dataset: monolingual wav2vec 2.0 large, XLS-R 53 (pre-trained on 53 languages), and XLS-R 128 (pre-trained on 128 languages). Table 3.7 presents these comparisons.

XLS-R 53 vs. Monolingual. XLS-R 53 shows generally worse performance than its monolingual counterpart. Overall accuracy drops from 93.17% to 90.07%, with 72.30% of subgroups showing decreased performance. The most severe degradation reaches −55.83% (t=11.49, upper part of Table 3.7), while the best improvement is 23.84% (t=6.39). Figure 3.9(a) shows this negative skew in the performance distribution.

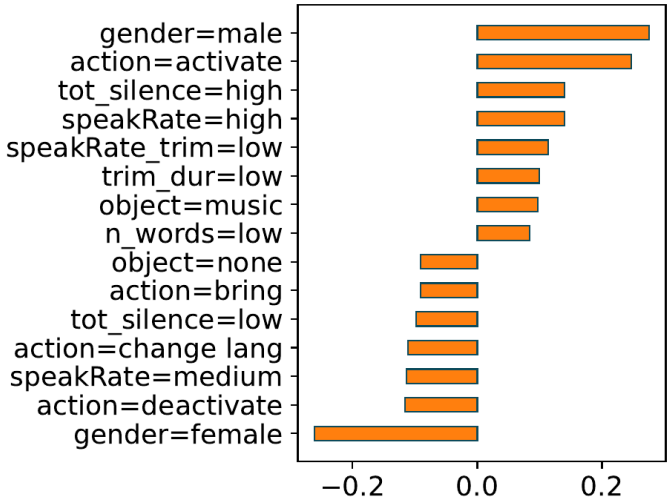


Fig. 3.10 **RQ4, FSC, Global Shapley values.** Accuracy gap for wav2vec 2.0 large monolingual to XLSR-128, top-15

XLS-R 128 vs. Monolingual. In contrast, XLS-R 128 demonstrates clear advantages over the monolingual model. Overall accuracy improves to 98.34%, with 92.90% of subgroups showing better performance. The largest improvement reaches 24.06% (t=6.14). Even in the small subset of subgroups (4.05%) showing decreased performance, the changes are minimal (-4.51%) and not statistically significant.

Attribute Impact Analysis. Global Shapley analysis (Figure 3.10) reveals interesting patterns in the transition to multilingual models. Gender effects are pronounced: female gender negatively impacts performance while male gender shows positive influence. Speaking characteristics like high speaking rate and silence duration show consistent positive impacts. Certain actions (e.g., "*activate*") and objects (e.g., "*music*") benefit particularly from multilingual training.

Summary of findings. Our study showed that switching from the monolingual model to XLS-R 53 on FSC does not improve performance, either overall or at the subgroup level. In contrast, using XLS-R 128 provides clear benefits, outperforming the monolingual model both overall and within subgroups. These results emphasize the importance of carefully evaluating each model for the specific task and dataset.

Final Remarks

Our study highlights the importance of analyzing speech model performance at the subgroup level across multiple tasks, datasets, and model variants. Performance disparities rarely arise from a single factor and often result from interactions between attributes, such as gender combined with speaking rate or emotion labels with acoustic conditions. Gender and speaking characteristics, particularly rate and duration, consistently influence model behavior. Increasing model size does not uniformly improve performance; gains vary by task and dataset, with ASR benefiting more consistently than emotion recognition. Model architecture affects subgroups differently: wav2vec 2.0 excels in emotion recognition, while HuBERT shows strengths in intent classification, but no architecture benefits all subgroups equally. Multilingual pretraining also shows nuanced effects: XLS-R 128 outperforms smaller or monolingual models overall, yet improvements are not consistent across all speaker groups. These results emphasize that aggregate metrics can hide substantial disparities and that responsible deployment of speech models requires careful evaluation of subgroup performance to ensure fairness and reliability.

Our approach enables detailed analysis of model performance and comparison at the subgroup level, providing several practical benefits. First, it allows practitioners to identify which subgroups gain or lose from a specific model, helping determine whether the model can be trusted. Second, it supports model debugging by highlighting subgroups affected by below-average performance. Third, practitioners can use these insights to actively improve the model for disadvantaged subgroups. Finally, the approach facilitates informed model selection, allowing choices to be guided by subgroup-level performance criteria rather than solely by overall metrics.

3.4 Post-Processing Mitigation via Divergence-Aware Data Acquisition

This section presents a targeted data acquisition strategy for mitigating performance disparities in already trained speech models, based on the work published in [21]. While acquiring additional training data is a common approach for improving model performance, indiscriminate data collection can be both costly and ineffective at addressing specific model weaknesses. We propose instead a focused strategy that

identifies and acquires data from subgroups where the model currently underperforms.

3.4.1 Motivation

Organizations constantly seek new data to improve their models' performance. This creates an important question: is it better to collect *all* available data, or focus on specific, carefully chosen subsets? Our research shows that selecting targeted subsets of data can improve models more effectively than collecting everything, while also reducing costs.

Consider a trained model M that performs differently across different groups of speakers. The traditional solution would be to collect more training data from all possible sources. However, this approach has important problems. Collecting and annotating large amounts of speech data is expensive. Much of this new data might come from groups where the model already works well. Adding data without careful selection might even make existing performance gaps worse.

Our approach uses the subgroup analysis framework presented in Section §3.3 to guide data collection. We first identify which groups of speakers the model struggles with most, focusing on those showing significant negative divergence. We then collect new data specifically for these challenging cases. This targeted approach makes better use of resources by collecting only the most helpful data. It directly addresses known model weaknesses by focusing on problematic cases. Because we use interpretable subgroups, we know exactly what kind of data we need to collect.

3.4.2 Methodology

Given a trained model M and its performance measure f, our data acquisition strategy proceeds in three steps.

Subgroup identification. Using the divergence analysis framework (§3.3), we identify the top-K subgroups with the largest negative divergence Δ_f. These represent the populations where the model most significantly underperforms.

Data acquisition. For each identified subgroup, we acquire new training samples that match the subgroup's characteristics. Because our subgroups are defined through

interpretable metadata, we can specifically target data collection to match these attributes.

Model updating. The model is retrained using both the original training data and the newly acquired samples, with the goal of improving performance on the previously underperforming subgroups while maintaining performance on others.

Formally, let T be the initial training set used to train model M, and let U be a set of additional utterances available for acquisition. Our goal is to identify a subset of U that, when added to T, will most effectively improve the model's performance on underperforming subgroups. We formalize this process as follows.

We first identify the set of challenging subgroups S^- as those showing negative divergence:

$$S^- = \{S \in \mathscr{S} | \Delta_f(S) < 0\} \tag{3.5}$$

where $\mathscr{S}$ is the set of all frequent subgroups and $\Delta_f(S)$ is the performance divergence as defined in Section §3.3.

To reduce redundancy among the challenging subgroups, we apply a pruning step following the approach outlined in [15]. When two overlapping subgroups S_a and S_b (where S_b includes S_a plus additional conditions) have similar divergence values, we retain only the more general subgroup S_a. This pruning yields a summarized set $\dot{S}^-$ of non-redundant challenging subgroups.

From this pruned set, we select the top-K subgroups with the highest absolute divergence, denoted as $\dot{S}^-_K$. The parameter K controls the breadth of our acquisition strategy. We then identify utterances in U that belong to these subgroups. An utterance $x_i \in U$ satisfies a subgroup S if its metadata values match all conditions defined by S, denoted as $x_i \models S$. The set of utterances to be acquired is thus:

$$U(\dot{S}^-_K) = \{x_i \in U | \exists S \in \dot{S}^-_K : x_i \models S\} \tag{3.6}$$

After acquiring new data, we retrain the model on the augmented dataset $T \cup U(\dot{S}^-_K)$. This retraining process aims to improve performance on underperforming subgroups while maintaining good performance on others. We use standard fine-tuning while monitoring performance on both the newly acquired data and the original training set. Training continues until the model converges on a validation set or reaches a maximum number of epochs.

Relationship to Other Acquisition Strategies

Our approach to data acquisition differs from existing methods in several important ways.

Random acquisition. Random sampling collects data without considering where models need improvement. Our method instead focuses specifically on data from groups where the model performs poorly.

Clustering-based approaches. Some methods, like [3], identify challenging cases by clustering speaker embeddings. While these clusters can find groups with similar problems, they don't tell us what makes these groups challenging. Our use of interpretable metadata makes it clear exactly what kind of data we need to collect.

Error-driven acquisition. Error-driven methods collect data from cases where the model makes mistakes. Our approach goes further by finding patterns in these errors. By identifying systematic problems rather than individual mistakes, we can collect data more efficiently.

Coverage-based approaches. The work of [123] also uses automatic subgroup identification through attribute combinations. However, they focus on finding groups that don't have enough representation in the training data. Our method instead looks for groups where the model performs poorly, regardless of how much data we already have. These approaches could work well together: their method ensuring we have enough data for each group, while ours targets specific performance problems. This combination would help create both well-represented and high-performing models.

3.4.3 Experimental Setup

To evaluate our divergence-aware data acquisition strategy, we design an experimental protocol[4] that allows us to assess both overall model improvements and subgroup-specific gains.

We partition each dataset into four distinct sets. The training set counts the 80% of original training data. The held-out set comprises the 20% left of original training data, and is used for data acquisition. The validation and test sets (original splits) are used for identifying problematic subgroups and final evaluation, respectively.

[4] github.com/koudounasalkis/Data-Acquisition-for-Speech-Model-Improvement

This organization allows us to simulate real-world data acquisition while maintaining consistent evaluation conditions. We first train the initial model on the training set, identify problematic subgroups using the validation set, acquire additional data from the held-out set, and finally evaluate on the test set.

Datasets and Models We evaluate our approach on two intent classification datasets: Fluent Speech Commands (FSC) and ITALIC (see Section §5.5 for more information). FSC provides a standard benchmark for English intent classification. ITALIC, which we introduced in [30], extends intent classification capabilities to Italian. It contains 16,521 samples from 70 speakers, with intents defined by action and scenario slots. Like FSC, ITALIC uses speaker-independent data splits to ensure robust evaluation.[5]

For model selection, we use architectures appropriate for each language. The English FSC dataset uses wav2vec 2.0 base, which contains 90M parameters. The Italian ITALIC dataset uses the multilingual XLS-R 128 model with 300M parameters. Both models use pre-trained checkpoints from the Hugging Face hub [149].

Metadata and evaluation metrics. We enrich both datasets with three types of metadata as described in Section §3.3: demographic information (gender, age, language fluency), acoustic characteristics (pause duration, word count, speaking rate), and task-specific features (intent components).

We evaluate performance using both overall and subgroup-specific metrics. Specifically, for overall metrics we employ accuracy and F1 Macro scores. For subgroup metrics, we opt for Δ^{-}_{max}, which measures the maximum negative divergence (worst-performing subgroup), Δ^{-}_{avg-n}, that measures the average divergence of top-N most divergent subgroups ($N = 10, 20, 50$), and $|\Delta_{avg-all}|$, which measures the average absolute divergence across all subgroups.

Baseline approaches. We compare our method against two baseline approaches for data acquisition.

Clustering-based acquisition. Following [3], we leverage acoustic embeddings to group similar speech samples together, resulting in 20 clusters for FSC and 10 for

[5] ITALIC appears only starting from this evaluation and not in our previous bias analysis as we developed it in parallel with those experiments. For detailed information about ITALIC metadata, see Section §5.5 or refer to the paper [30].

ITALIC.[6] We then identify which clusters show the worst performance and acquire similar samples from the held-out set. While this automated approach works well, it operates on abstract embeddings rather than interpretable characteristics, making it harder to understand what data we need to collect.

Random acquisition. We implement random sampling in two ways. The first matches the number of samples acquired by our method. The second matches the number acquired by the clustering baseline. This helps us evaluate whether improvements come from better sample selection or simply from adding more data.

Experimental protocol. We evaluate all approaches under two conditions. The *equal-data setting* ensures all methods acquire the same number of samples, enabling direct comparison. The *unrestricted setting* allows our method and the clustering approach to acquire more samples when they find additional problematic cases.

Our method acquires all held-out samples that match the metadata of critical subgroups. We control the breadth of acquisition through parameter K, testing values from 2 to 5. We also include results when acquiring the entire held-out set as an upper bound reference.

3.4.4 Results and Discussion

Tables 3.8 and 3.9 present the comparative evaluation of our divergence-aware acquisition strategy against baseline approaches. We analyze the results from multiple perspectives: overall performance gains, subgroup-specific improvements, and the impact of acquisition breadth.

Impact of targeted data acquisition. Our approach consistently outperforms both random and clustering-based baselines across all metrics. The highest improvements occur when targeting the top-2 most challenging subgroups ($K = 2$). For FSC, this configuration achieves improvement in F1 Macro (from 86.34% to 94.71%), reduction in maximum negative divergence (Δ^{-}_{max}) from -70.09% to -40.60%, and substantial reductions in average divergence across top-10, 20, and 50 subgroups.

[6] These values were found to be optimal through empirical testing. See [30] and project repository for details.

Table 3.8 **FSC, wav2vec 2.0 base.** Mean±std over three runs. We compare the *original* training, two baselines (*random*, *clustering*), and our *divergence-aware* acquisition in two scenarios: *equal-data* and *unrestricted*. *#Samples* is the number of added utterances from the held-out set. Δ^{-}_{max} is the most negative divergence, Δ^{-}_{avg-x} the average divergence over the top-x negative subgroups, and $\Delta_{avg-all}$ the average over all subgroups. Best per-K results in **bold**; overall best in light blue.

K	Approach	#Samples	Accuracy	F1 Macro	Δ^{-}_{max}	Δ^{-}_{avg-10}	Δ^{-}_{avg-20}	Δ^{-}_{avg-50}	$\lvert\Delta_{avg-all}\rvert$
-	original	-	91.58±0.08	86.34±0.13	-70.09±0.26	-70.09±0.26	-65.73±0.49	-53.31±0.19	1.06±0.07
				Equal-data setting					
2	random	226	92.56±0.44	90.25±0.60	-52.20±2.57	-51.11±2.19	-46.61±1.34	-43.98±0.68	0.97±0.02
	clustering	226	89.77±0.88	87.02±0.15	-47.37±0.42	-47.34±0.42	-47.23±0.43	-46.75±0.91	0.94±0.04
	ours	226	**96.55±0.08**	**94.71±0.12**	**-40.60±0.35**	**-40.28±0.36**	**-38.08±0.36**	**-32.72±0.28**	**0.81±0.03**
3	random	382	94.13±0.58	91.51±0.82	-52.99±3.40	-51.92±3.02	-49.39±2.21	-45.98±1.78	0.33±0.04
	clustering	382	90.03±0.97	85.30±0.94	-46.40±0.36	-45.02±0.33	-41.59±0.28	-37.79±0.16	0.81±0.02
	ours	382	93.62±0.29	**92.96±0.46**	**-42.23±0.12**	**-42.21±0.11**	**-41.48±0.11**	**-33.61±0.07**	**0.22±0.02**
4	random	422	92.64±0.27	91.29±0.21	-55.83±2.11	-55.71±2.04	-51.41±1.74	-45.41±1.74	0.39±0.02
	clustering	422	87.72±0.71	83.42±0.48	-47.59±0.25	-46.98±0.21	-45.69±0.12	-43.98±0.09	0.72±0.03
	ours	422	**95.16±0.11**	**92.47±0.22**	**-45.68±0.24**	**-44.56±0.25**	**-41.53±0.24**	**-37.02±0.20**	**0.15±0.01**
5	random	509	91.48±0.55	90.27±0.49	-54.82±3.41	-54.75±3.29	-54.69±3.11	-51.12±2.25	0.96±0.08
	clustering	509	91.44±1.41	87.92±1.38	-51.92±0.19	-51.90±0.24	-49.79±0.18	-43.39±0.11	0.45±0.03
	ours	509	**94.12±0.13**	**92.57±0.16**	**-49.33±0.15**	**-49.29±0.12**	**-48.11±0.21**	**-39.01±0.11**	**0.11±0.02**
				Unrestricted setting					
2	random	406	94.26±0.27	91.17±0.86	-54.26±1.14	-53.93±1.17	-53.24±1.12	-52.37±0.55	0.86±0.06
	clustering	406	92.94±0.07	90.82±1.19	-51.81±0.86	-51.22±0.92	-49.99±0.10	-48.52±0.11	1.24±0.09
	ours	226	**96.55±0.08**	**94.71±0.12**	**-40.60±0.35**	**-40.28±0.36**	**-38.08±0.36**	**-32.72±0.28**	**0.81±0.03**
3	random	874	92.21±0.49	90.30±0.55	-64.72±3.07	-62.10±2.49	-56.56±2.32	-51.57±1.88	0.38±0.06
	clustering	874	**94.47±0.44**	92.23±0.45	-47.33±0.33	-45.83±0.16	-42.73±0.21	-39.72±0.49	0.32±0.03
	ours	382	93.62±0.29	**92.96±0.46**	**-42.23±0.12**	**-42.21±0.11**	**-41.48±0.11**	**-33.61±0.07**	**0.22±0.02**
4	random	1046	91.31±0.98	89.48±0.52	-61.85±1.58	-60.72±1.28	-58.08±0.86	-54.83±1.00	1.19±0.03
	clustering	1046	93.28±0.19	91.42±0.18	-52.28±0.63	-51.08±0.58	-48.65±0.40	-45.35±0.44	0.85±0.09
	ours	422	**95.16±0.11**	**92.47±0.22**	**-45.68±0.24**	**-44.56±0.25**	**-41.53±0.24**	**-37.02±0.20**	**0.15±0.01**
5	random	1276	92.01±0.49	91.00±0.65	-67.77±1.96	-66.94±1.55	-65.31±1.23	-62.65±1.19	0.48±0.03
	clustering	1276	92.75±0.21	90.66±0.22	-61.04±0.19	-60.84±0.24	-57.84±0.18	-49.72±0.11	1.33±0.01
	ours	509	**94.12±0.13**	**92.57±0.16**	**-49.33±0.15**	**-49.29±0.12**	**-48.11±0.21**	**-39.01±0.11**	**0.11±0.02**
-	all data	4606	93.42±0.17	93.11±0.17	-53.18±0.15	-50.89±0.09	-45.61±0.14	-40.37±0.16	0.37±0.01

Similar patterns emerge for ITALIC, where our $K = 2$ configuration yields an F1 Macro improvement from 68.08% to 72.51%, reduction in Δ^{-}_{max} from -47.63% to -31.75%, and consistent improvements in subgroup-level metrics.

Notably, these improvements are achieved with significantly less additional data than required by the clustering-based approach. For example, on FSC, our method consistently outperforms the clustering baseline while using less than half the number of added samples across almost all values of K. This highlights the efficiency of targeted data acquisition in improving model robustness without the cost of extensive acquisition.

Table 3.9 **ITALIC, XLS-R 128 large.** Mean±std over three runs. Best per-K results in **bold**; overall best in light blue.

K	Approach	#Samples	Accuracy	F1 Macro	Δ^{-}_{max}	Δ^{-}_{avg-10}	Δ^{-}_{avg-20}	Δ^{-}_{avg-50}	$\lvert\Delta_{avg-all}\rvert$
-	original	-	73.79±0.32	68.08±0.37	-47.63±1.93	-47.52±1.94	-47.15±1.92	-43.31±1.78	0.60±0.01
				Equal-data setting					
2	random	154	75.32±0.63	70.72±0.58	-47.00±0.81	-46.86±0.80	-46.22±0.77	-41.68±0.70	0.38±0.02
	clustering	154	74.05±0.33	69.09±0.75	-45.02±2.02	-44.91±2.01	-44.14±1.81	-39.79±1.33	0.37±0.08
	ours	154	**77.40±0.24**	**72.51±0.14**	**-31.75±0.55**	**-31.71±0.55**	**-31.11±0.41**	**-28.19±0.18**	0.34±0.03
3	random	252	75.81±0.13	71.46±0.25	-51.32±1.30	-51.14±1.29	-50.27±1.16	-45.37±0.89	0.25±0.01
	clustering	252	75.87±0.20	70.70±0.31	-42.93±0.52	-42.82±0.51	-41.89±0.51	-37.25±0.48	0.25±0.02
	ours	252	**76.50±0.30**	**71.69±0.59**	**-36.73±0.33**	**-36.57±0.32**	**-36.18±0.30**	**-32.20±0.57**	**0.17±0.03**
4	random	540	75.67±0.20	71.71±0.02	-41.07±0.69	-40.96±0.68	-40.36±0.72	-36.41±0.77	0.34±0.05
	clustering	540	75.76±0.21	71.22±0.17	-41.50±0.80	-41.40±0.80	-40.79±0.80	-37.87±0.71	0.22±0.03
	ours	540	**76.29±0.13**	**72.48±0.48**	**-37.30±1.05**	**-37.22±1.04**	**-36.79±1.03**	**-33.42±0.75**	**0.16±0.04**
5	random	604	75.13±0.05	71.26±0.17	-41.91±1.95	-41.79±1.94	-41.09±1.83	-37.53±1.26	0.29±0.04
	clustering	604	75.88±0.29	70.60±0.59	-40.41±1.17	-40.32±1.16	-39.47±1.10	-36.12±0.81	0.19±0.03
	ours	604	77.14±0.04	71.32±0.41	**-37.52±0.78**	**-37.44±0.77**	**-36.83±0.76**	**-33.75±0.40**	**0.09±0.02**
				Unrestricted setting					
2	random	383	75.34±0.32	69.75±0.59	-40.12±1.47	-40.01±1.46	-39.21±1.33	-35.81±0.84	0.37±0.03
	clustering	383	74.35±0.12	69.51±0.30	-41.64±0.60	-41.52±0.60	-40.84±0.52	-36.90±0.38	**0.32±0.02**
	ours	154	**77.40±0.24**	**72.51±0.14**	**-31.75±0.55**	**-31.71±0.55**	**-31.11±0.41**	**-28.19±0.18**	0.34±0.03
3	random	548	76.38±0.12	71.09±0.43	-40.51±1.07	-40.41±1.06	-39.52±1.03	-35.12±0.80	0.23±0.04
	clustering	548	75.71±0.12	71.31±0.18	-39.74±2.24	-39.65±2.22	-38.81±2.09	-35.02±1.70	0.29±0.01
	ours	252	**76.50±0.30**	**71.69±0.59**	**-36.73±0.33**	**-36.57±0.32**	**-36.18±0.30**	**-32.20±0.57**	**0.17±0.03**
4	random	945	75.90±0.30	70.83±0.39	-42.34±1.23	-42.23±1.22	-41.65±1.15	-37.83±0.76	0.19±0.02
	clustering	945	76.02±0.40	71.53±0.63	-42.52±3.26	-42.43±3.24	-41.76±3.16	-37.97±2.71	0.26±0.06
	ours	540	**76.29±0.13**	**72.48±0.48**	**-37.30±1.05**	**-37.22±1.04**	**-36.79±1.03**	**-33.42±0.75**	**0.16±0.04**
5	random	1035	**77.19±0.34**	71.51±0.39	-46.37±0.88	-46.27±0.89	-45.73±0.98	-41.38±1.11	0.21±0.05
	clustering	1035	77.05±0.22	**71.93±0.04**	-42.93±1.04	-42.86±1.05	-42.39±1.11	-38.34±1.14	0.18±0.03
	ours	604	77.14±0.04	71.32±0.41	**-37.52±0.78**	**-37.44±0.77**	**-36.83±0.76**	**-33.75±0.40**	**0.09±0.02**
-	all data	2625	75.71±0.36	73.22±0.33	-47.54±0.79	-47.36±0.76	-46.68±0.47	-41.93±0.00	0.15±0.03

Effect of acquisition breadth. The parameter K controls how many challenging subgroups we target for data acquisition. Our analysis reveals an interesting trade-off.

K=2 (Focused acquisition). This configuration achieves the best overall performance in terms of accuracy and F1 Macro, and shows the largest improvement in worst-case performance (Δ^{-}_{max}), all while requiring minimal additional data.

Larger K values. On the contrary, this configuration leads to a slight decrease in overall performance compared to $K = 2$, but also to a more uniform improvement across subgroups. $K = 5$ achieves the lowest average absolute divergence ($|\Delta_{avg-all}|$).

This pattern suggests that focused acquisition ($K = 2$) is most effective for addressing severe performance disparities, while broader acquisition helps achieve more balanced improvements across all subgroups.

Comparison with complete data acquisition. A particularly noteworthy finding is that our targeted approach outperforms acquiring the entire held-out set. On FSC,

our $K = 2$ configuration achieves 94.71% F1 Macro vs. 93.11% with all data. On ITALIC, we achieve 72.51% F1 Macro with $K = 2$ vs 73.22% with all data, but with significantly better subgroup performance.

This demonstrates that strategic data selection can be more effective than indiscriminate data acquisition, while requiring only a fraction of the data and computational resources.

Analysis of performance patterns. Our analysis reveals several key patterns in how targeted acquisition affects different types of subgroups.

Cross-group performance gap. Our targeted acquisition strategy shows particular effectiveness in reducing performance disparities between demographic groups. For FSC, before acquisition, we observe a gap in accuracy between female (93.5%) and male (89.7%) speakers. After applying our approach with $K = 2$, this disparity decreases, with female speakers achieving 97.8% accuracy and male speakers reaching 95.3%. This represents a reduction in the gender performance gap from 3.8% to 2.5%, demonstrating that our targeted acquisition strategy not only improves overall performance but also helps create a more equitable model.

Speaking characteristics. Analysis of subgroups defined by speaking characteristics reveals a consistent pattern of improvement. For instance, subgroups characterized by high speaking rates and short durations, which initially showed some of the largest negative divergences, see substantial improvements after targeted acquisition. This suggests our approach effectively addresses variability in speaking patterns that often challenge speech models.

3.4.5 Summary and Practical Implications

The experimental results demonstrate several important advantages for real-world deployment of speech models. Our targeted approach achieves superior performance while acquiring only 5-10% of the available data. This selective strategy significantly reduces costs associated with data collection and annotation.

The ability to identify and focus on specific challenging subgroups enables precise improvements where they are most needed. Rather than collecting data indiscriminately, organizations can address known performance disparities directly. This

targeted approach proves particularly valuable when resources for data collection are limited.

The strong performance with small K values, especially $K = 2$, shows that highly focused acquisition strategies can be remarkably effective. Organizations do not need to collect large amounts of additional data to see significant improvements. However, the choice of K offers important flexibility in deployment. Focused acquisition ($K = 2$) maximizes performance improvements for the most problematic cases. Broader acquisition ($K = 5$) achieves more uniform performance across all subgroups. This trade-off allows organizations to align their data collection strategy with specific application needs and fairness requirements.

3.5 In-Processing Mitigation Strategies

While the post-processing data acquisition strategy presented in Section §3.4 can effectively improve model performance, addressing biases during the training process itself often proves more effective and computationally efficient. As demonstrated in our paper [20], incorporating bias mitigation directly into the training pipeline allows for more systematic improvements in model fairness.

This section presents two complementary in-processing approaches for mitigating subgroup performance disparities introduced in [20]: divergence-aware regularization (§3.5.2) and targeted data augmentation (§3.5.3). Both strategies leverage the subgroup analysis framework presented in §3.3, but differ in how they use this information to guide model training.

3.5.1 Problem Formulation

Given a model M being trained on dataset D, our goal is to reduce performance disparities across subgroups while maintaining or improving overall performance. Let S^- be the set of challenging, i.e., problematic subgroups identified through our divergence analysis:

$$S^- = \{S \in \mathscr{S} | \Delta_f(S) < 0\} \tag{3.7}$$

Unlike the post-processing approach that requires acquiring additional data, in-processing mitigation operates on the existing training set by either modifying the

loss function to focus more on samples from underperforming subgroups, or by augmenting existing samples belonging to challenging subgroups to improve model robustness.

3.5.2 Divergence-Aware Regularization

We propose a divergence-based regularization term to mitigate subgroup disparities during training. The key insight is that the higher the divergence of a subgroup, the more attention the model should pay to samples from that subgroup during training. We implement this through a sample weighting mechanism where samples from subgroups with higher divergence receive greater weights in the loss computation.

Let T and V be the training and validation sets, respectively. For a model M_e at epoch e trained on T, we first extract the set S of frequent subgroups and their divergence scores from the validation set V. For each utterance x_i in T with true label y_i and predicted label $\hat{y}_i$, we define $\mathscr{S}(x_i)$ as the set of subgroups satisfied by x_i:

$$\mathscr{S}(x_i) = \{S \in \mathscr{S} \mid x_i \vdash S\} \tag{3.8}$$

We then assign each training utterance a boosting weight based on the maximum absolute divergence across its subgroups:

$$w(x_i) = \max_{S \in \mathscr{S}(x_i)} |\Delta_f(S, M)| \tag{3.9}$$

Using these weights, we define our divergence-based regularization loss $\mathscr{L}_\Delta$:

$$\mathscr{L}_\Delta = \sum_{x_i \in T} w(x_i) \mathscr{L}_{CE}(y_i, \hat{y}_i) \tag{3.10}$$

where $\mathscr{L}_{CE}$ denotes the standard cross-entropy loss.[7] The final training objective combines this regularization term with the standard loss:

$$\mathscr{L} = \alpha \mathscr{L}_{CE} + (1 - \alpha) \mathscr{L}_\Delta \tag{3.11}$$

where α controls the strength of the regularization.

[7] Since the task is standard intent classification, we use the cross-entropy loss. This choice can be adapted according to the specific task; for instance, ASR tasks would typically employ the CTC loss instead.

Algorithm 1 Divergence-Aware Regularization

Require: Training Set T, Validation Set V, min frequency u
Ensure: M: Model
1: Initialize weights $w(x_i) = 1.0 \; \forall x_i \in T$
2: $T_m, V_m \leftarrow$ Derive metadata T, V
3: **for each** epoch $e \in E$ **do**
4: $M_e \leftarrow$ Model trained on T at epoch e via Eq. 3.11
5: $S, \Delta_f(S, M_e) \forall S \in S \leftarrow$ DIVEXPLORER(M_e, V_m)
6: $S(x_i) \leftarrow$ Satisfy$(x_i, S) \; \forall x_i \in T_m$
7: $w(x_i) \leftarrow$ ComputeWeights$(x_i, S(x_i)) \; \forall x_i \in T_m$ via Eq. 3.9
8: **end for**
9: **return** M_e

Algorithm 1 provides a complete overview of the divergence-aware regularization process. The algorithm begins by initializing unit weights for all training samples and deriving metadata for both training and validation sets (Line 2). For each training epoch, it proceeds through four main steps. First, it trains the model using the combined loss from Equation 3.11. Second, it uses DivExplorer to extract subgroups and their divergence scores from the validation set (Line 5). Third, it identifies which subgroups each training sample belongs to (Line 6). Finally, it updates the sample weights based on the maximum divergence of their corresponding subgroups using Equation 3.9 (Line 7). This iterative process ensures that the model continuously adapts its focus toward the most challenging subgroups. Finally, the algorithm produces the final boosted model as output (Line 9).

3.5.3 Divergence-Aware Data Augmentation

While regularization modifies the training process through loss weighting, our second strategy directly modifies training data to help models learn better from challenging cases. At each training epoch, we identify subgroups where the model performs worst and augment samples from these groups. This targeted augmentation helps the model improve its performance on historically underperforming subgroups.

Let V be the validation set and M_e be the model at epoch e. We first identify the top-K challenging subgroups $\dot{S}^-$ showing the largest absolute divergence on V. These subgroups represent where model M_e most needs improvement. We obtain $\dot{S}^-$ using the same pruning process described in our post-processing approach (§3.4).

For each training batch T_b, we identify samples belonging to challenging subgroups:

$$T_b(\dot{S}_K^-) = \{x_i \in T_b \mid \exists S \in \dot{S}_K^- : x_i \vdash S\} \tag{3.12}$$

This formula follows the same logic as our post-processing acquisition, but operates on training data rather than new samples to acquire. We then augment these identified samples using various techniques. These include changing the speech rate through time stretching, adding background noise, applying reverberation effects, shifting the pitch, or combining multiple transformations.

The model then trains on this augmented batch. By providing more variations of problematic cases during training, the model learns to handle difficult samples more effectively. This improves its performance on previously underperforming groups while maintaining good performance on others.

3.5.4 Implementation Considerations

Both mitigation strategies require careful implementation choices to work effectively.

Regularization parameters. The weighting parameter α in our divergence-aware regularization balances overall performance with subgroup improvements. Our experiments show that $\alpha = 0.7$ works best, providing enough focus on challenging subgroups while keeping training stable. Lower values make training unstable by focusing too much on difficult samples. Higher values reduce the effect of our mitigation strategy.

Data augmentation strategy. We adapt our augmentation techniques based on what makes each subgroup challenging. For groups that struggle with speaking rate, we apply time stretching between 0.8 and 1.2 times the original speed. When acoustic robustness is the issue, we add background noise at 5-15dB signal-to-noise ratio (SNR) levels. For speaker variations, we shift pitch within 20% of the original. Environmental effects are simulated through room impulse response convolution. We apply these transformations randomly during training, with probabilities proportional to the subgroups divergence.

Computational considerations. Both strategies introduce additional overhead during training, but in different ways. The regularization approach adds minimal computation, requiring only divergence calculations once per epoch and efficient

sample weight updates. Data augmentation can be more demanding, depending on the transformations used.

Relationship to post-processing mitigation. While both in-processing and post-processing approaches address subgroup disparities, they offer different advantages. In-processing strategies work with existing training data, avoiding the need for additional data collection. They address disparities more systematically by modifying the learning process itself. These methods often achieve better parameter efficiency by optimizing directly for fairness during training. However, they cannot address issues caused by fundamental gaps in the training data. They may also need longer training times and more careful tuning to maintain stability. The choice between approaches depends on practical constraints like data availability, computational resources, and application needs. Both methods can work together effectively when circumstances allow.

3.5.5 Experimental Setup

Our experimental evaluation[8] spans three tasks: intent classification, speech emotion recognition, and automatic speech recognition. For these tasks, we select four datasets that cover both English and Italian languages: FSC and ITALIC for IC, IEMOCAP for SER, LibriSpeech "clean-360" for ASR.

We implement two distinct experimental configurations to evaluate our mitigation techniques and enable a direct comparison between in-processing and post-processing approaches. The first configuration uses the full training set for model training, employs the validation set to identify frequent and challenging subgroups, and uses the test set for final evaluation. This setup is used exclusively to assess in-processing techniques. The second configuration splits the training set, using 80% for training and reserving 20% for post-processing data acquisition (as described in Section §3.4). This split preserves speaker independence, ensuring that no speaker appears in more than one set.

For model architecture selection, we fine-tune pre-trained checkpoints from the Hugging Face hub [149]. We use wav2vec 2.0 base for FSC and IEMOCAP tasks, XLS-R 128 for ITALIC, and Whisper base monolingual for LibriSpeech.

[8] github.com/koudounasalkis/Divergence-Aware-Dual-Strategy-Mitigation

We set minimum support threshold to $u = 0.03$. For data acquisition and augmentation strategies, we explore K values from 2 to 5, with $K = 2$ serving as our primary setting based on empirical results on data acquisition [21].

We evaluate performance using task-appropriate metrics. For IC and ER tasks, we measured accuracy and macro F1 scores. ASR performance is assessed using WER and Character Error Rate (CER). We also track subgroup-specific metrics including maximum negative divergence ($\Delta^{-}max$), average divergence across top-N subgroups (Δ^{-}_{avg-N}), and average absolute divergence ($|\Delta^{-}_{avg-all}|$).

Baselines. We evaluate our mitigation approaches against four established baseline methods for identifying challenging samples.

The first baseline uses random selection of problematic samples. This simple approach helps demonstrate the advantages of more sophisticated selection strategies.

The second baseline follows the cluster-based approach proposed in [3]. We test this approach with two types of embeddings. The first variant (`clustering`) uses acoustic embeddings from the last hidden state of the models. The second variant ($\texttt{clustering}_{\texttt{X}}$) employs x-vector speaker embeddings [57], which have shown strong results for ASR tasks [3]. For clustering, we apply K-means with dataset-specific cluster numbers: 50 for LibriSpeech, 20 for FSC and IEMOCAP, and 10 for ITALIC. These numbers were chosen based on empirical performance optimization [21]. After clustering, we identify the clusters with poorest performance and select challenging samples based on proximity to these underperforming clusters.

The third baseline uses K-Nearest Neighbors (KNN) to identify challenging samples. This approach classifies an utterance as challenging based on majority voting among its neighbors in the validation set. We optimize K values through validation set performance: 14 for FSC, 11 for ITALIC, 12 for IEMOCAP, and 18 for LibriSpeech.

The fourth baseline implements an error-driven approach similar to Magar et al. [155]. This method identifies samples that the model predicts incorrectly in the held-out set and labels them as problematic. We apply this error-driven approach only for post-processing mitigation, as error handling is already part of standard training loss functions. This final baseline requires ground truth labels for the held-out set, which may not always be available in practice.

Table 3.10 **IC, Mean±std over three runs.** Results for the original fine-tuning and mitigation strategies, including targeted data acquisition (`data acq.`), data augmentation (`data++`), and regularization (`regular.`), using the original training set split into training and held-out sets, with $K = 2$. The analyzed selection methods are Clustering (`Clu` and $\texttt{Clu}_\texttt{X}$), Random (`Rand`), and Error-Driven (`Err`). Best results per dataset are shown in **bold**, second-best are underlined, and the best per dataset and strategy are highlighted in light blue.

DS	*Approach*	*Strategy*	*Accuracy*	*F1 Macro*	Δ^{-}_{max}	Δ^{-}_{avg-10}	Δ^{-}_{avg-20}	Δ^{-}_{avg-50}	$\lvert\Delta_{avg-all}\rvert$
FSC	original	-	91.58±0.08	86.34±0.13	-70.09±0.26	-70.09±0.26	-65.73±0.49	-53.31±0.19	1.06±0.07
	Rand	data acq.	92.56±0.44	90.25±0.60	-52.20±2.57	-51.11±2.19	-46.61±1.34	-43.98±0.68	0.97±0.02
	KNN	data acq.	92.07±0.17	89.92±0.11	-49.90±0.33	-49.85±0.29	-49.76±0.27	-46.98±0.28	0.96±0.03
	Clu	data acq.	89.77±0.88	87.02±0.15	-47.37±0.42	-47.34±0.42	-47.23±0.43	-46.75±0.91	0.94±0.04
	$\texttt{Clu}_\texttt{X}$	data acq.	91.44±0.65	90.12±0.66	-47.99±0.53	-47.95±0.52	-47.80±0.49	-47.11±0.44	0.89±0.05
	Err	data acq.	95.71±0.74	94.06±0.83	-48.13±0.39	-48.02±0.36	-47.58±0.35	-45.97±0.48	0.92±0.04
	ours	data acq.	**96.55±0.08**	94.71±0.12	-40.60±0.35	-40.28±0.36	-38.08±0.36	-32.72±0.28	0.81±0.03
	Rand	data++	92.85±0.75	92.29±0.68	-45.67±2.78	-45.59±2.75	-43.41±2.68	-41.28±2.51	0.84±0.27
	KNN	data++	93.94±0.28	93.15±0.31	-43.61±1.32	-43.34±1.24	-42.12±1.19	-38.84±1.08	0.75±0.03
	Clu	data++	94.49±0.41	94.31±0.44	-40.09±2.12	-39.95±2.03	-39.77±1.84	-34.65±1.07	0.38±0.10
	$\texttt{Clu}_\texttt{X}$	data++	95.12±0.44	95.02±0.45	-41.13±1.89	-41.01±1.85	-40.45±1.74	-39.89±1.61	0.37±0.06
	ours	data++	95.75±0.37	95.48±0.35	-36.12±0.39	-35.98±0.37	-34.77±0.36	-32.65±0.33	0.35±0.04
	Rand	regular.	93.41±0.52	93.22±0.67	-44.51±6.59	-44.25±6.55	-44.04±6.21	-38.54±5.85	0.85±0.14
	KNN	regular.	95.11±0.21	95.04±0.20	-41.32±3.52	-41.19±3.28	-40.51±3.15	-36.95±2.75	0.62±0.05
	Clu	regular.	95.75±0.39	95.49±0.41	-39.51±5.68	-39.18±5.21	-37.29±4.74	-34.74±4.18	0.43±0.02
	$\texttt{Clu}_\texttt{X}$	regular.	96.04±0.38	95.99±0.36	-39.88±4.17	-39.80±4.14	-38.71±4.03	-36.13±3.98	0.38±0.03
	ours	regular.	96.47±0.11	**96.33±0.12**	**-34.49±0.45**	**-34.49±0.45**	**-34.11±0.41**	**-31.34±0.32**	**0.29±0.01**
	original	all data	93.42±0.17	93.11±0.17	-53.18±0.15	-50.89±0.09	-45.61±0.14	-40.37±0.16	0.37±0.01
ITALIC	original	-	73.79±0.32	68.08±0.37	-47.63±1.93	-47.52±1.94	-47.15±1.92	-43.31±1.78	0.60±0.01
	Rand	data acq.	75.32±0.63	70.72±0.58	-47.00±0.81	-46.86±0.80	-46.22±0.77	-42.68±0.70	0.48±0.02
	KNN	data acq.	75.56±0.57	70.21±0.54	-46.11±0.93	-46.02±0.92	-45.49±0.84	-42.17±0.74	0.39±0.02
	Clu	data acq.	74.05±0.33	69.09±0.75	-45.02±2.02	-44.91±2.01	-44.14±1.81	-39.79±1.33	0.37±0.08
	$\texttt{Clu}_\texttt{X}$	data acq.	76.19±0.37	71.04±0.64	-46.51±1.87	-46.48±1.85	-45.77±1.63	-42.48±1.29	0.37±0.01
	Err	data acq.	77.14±0.52	72.65±0.63	-46.97±1.15	-46.84±1.07	-45.91±1.02	-42.36±0.93	0.45±0.04
	ours	data acq.	**77.40±0.24**	72.51±0.14	-31.75±0.55	-31.71±0.55	-31.11±0.41	-28.19±0.18	0.34±0.03
	Rand	data++	75.14±0.49	73.01±0.79	-46.89±2.05	-46.51±1.98	-44.98±1.57	-42.04±1.36	0.35±0.12
	KNN	data++	75.97±0.34	73.67±0.39	-41.19±1.17	-40.53±1.06	-38.57±0.95	-35.77±0.89	0.31±0.03
	Clu	data++	76.59±0.84	73.98±0.78	-38.95±2.69	-38.37±2.43	-37.01±2.20	-34.15±2.02	0.28±0.04
	$\texttt{Clu}_\texttt{X}$	data++	76.94±0.65	74.01±0.67	-39.62±2.29	-39.57±2.25	-38.43±2.11	-35.04±1.87	0.25±0.03
	ours	data++	77.12±0.54	74.05±0.42	-31.93±1.91	-31.58±1.85	-30.05±1.59	-28.19±1.35	0.23±0.05
	Rand	regular.	76.04±0.71	72.11±0.55	-46.58±2.29	-46.22±2.21	-45.87±2.08	-43.16±1.97	0.33±0.11
	KNN	regular.	76.54±0.44	73.08±0.39	-41.23±1.24	-41.04±1.17	-38.63±1.02	-35.78±0.87	0.29±0.04
	Clu	regular.	76.67±0.79	74.01±0.76	-38.43±2.51	-38.05±2.18	-36.59±1.96	-33.93±1.79	0.25±0.03
	$\texttt{Clu}_\texttt{X}$	regular.	76.98±0.58	74.16±0.53	-39.15±2.21	-39.11±2.17	-37.89±2.05	-34.14±1.85	0.24±0.03
	ours	regular.	77.02±0.61	**74.19±0.48**	**-31.54±2.02**	**-31.14±1.93**	**-29.88±1.74**	**-28.10±1.67**	0.21±0.05
	original	all data	75.71±0.36	73.22±0.33	-47.54±0.79	-47.36±0.76	-46.68±0.47	-41.93±0.00	**0.15±0.03**

3.5.6 Results and Discussion

Our experimental evaluation examines the effectiveness of our proposed mitigation strategies compared to existing approaches. We present results in two configurations: using a held-out portion of the training data, and using the complete training set. This evaluation allows us to assess both in-processing and post-processing methods under comparable conditions.

Performance Analysis

Baseline comparison. Tables 3.10 and 3.11 present our mitigation results compared to baseline approaches. Our strategies consistently outperform all baselines across all datasets. For IC and SER tasks, we achieve higher accuracy and F1 scores, and for ASR tasks, we obtain lower WER and CER values. The improvements are particularly notable in subgroup-specific metrics, with significant reductions in maximum negative divergence (Δ^-_{max}). We also achieve better average divergence scores across top-10, top-20, and top-50 underperforming subgroups (Δ^-_{avg-n}). The average absolute divergence across all groups ($|\Delta^-_{avg-all}|$) shows consistent improvement.

In post-processing scenarios (i.e., data acquisition, see §3.4), the error-driven baseline ranks second for overall performance. However, for subgroup-specific improvements, the clustering-based approach emerges as the second-best method. For in-processing techniques, clustering-based approaches consistently rank second in both overall and subgroup-level metrics. Interestingly, the speaker embedding variant ($\texttt{clustering}_{\texttt{X}}$) shows better overall accuracy but lower subgroup-level performance.

In-processing vs. post-processing analysis. Our analysis reveals that addressing performance disparities during training proves more effective than post-processing modifications. Tables 3.10 and 3.11 demonstrate consistently lower divergence scores for in-processing methods.

Both divergence-aware regularization and targeted data augmentation achieve superior results compared to post-processing data acquisition. Among in-processing techniques, divergence-aware regularization emerges as the most effective approach. This method shows superior performance in both overall metrics and subgroup-level evaluations.

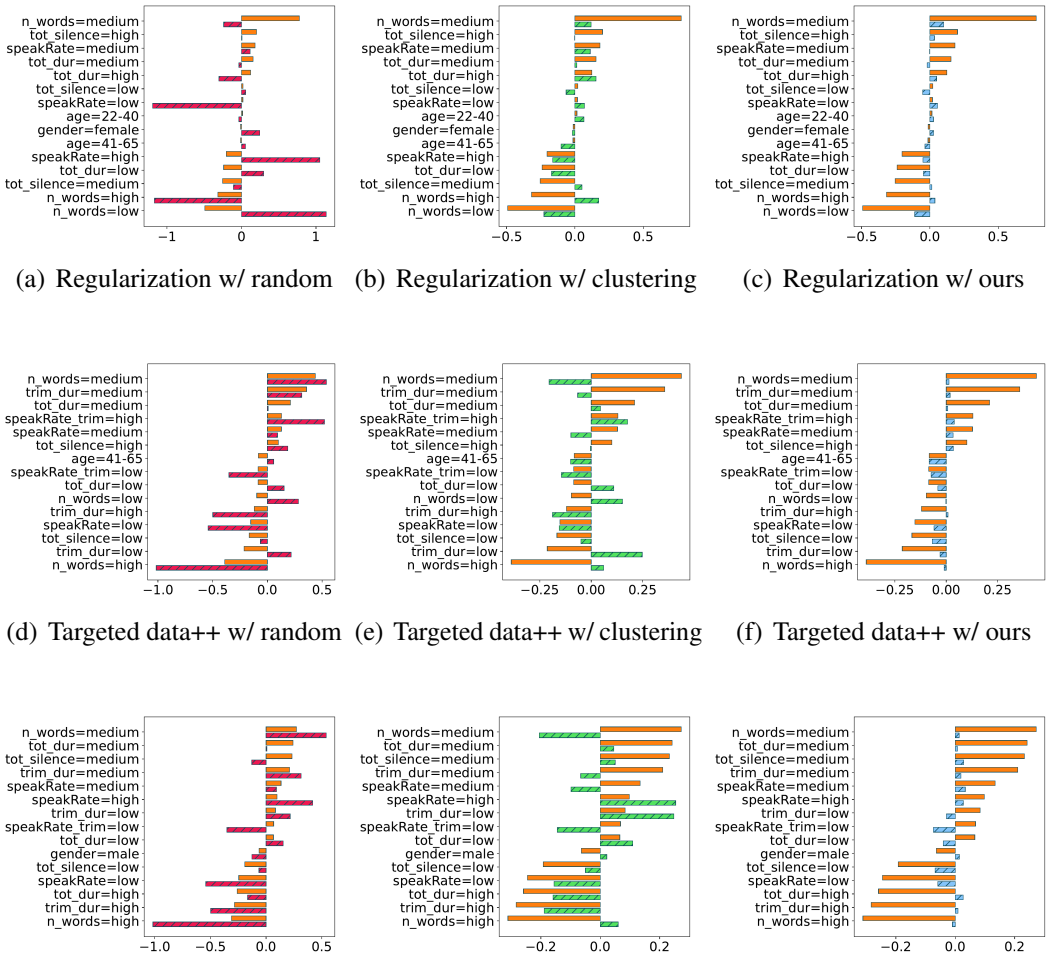
Table 3.11 **Mean±std over three runs on IEMOCAP and LibriSpeech datasets.** Best results for each dataset are shown in **bold**, second-best underlined, and the top result for each dataset and mitigation strategy highlighted in light blue.

DS	*Approach*	*Strategy*	*Accuracy*	*F1 Macro*	Δ^{-}_{max}	Δ^{-}_{avg-10}	Δ^{-}_{avg-20}	Δ^{-}_{avg-50}	$\lvert\Delta_{avg-all}\rvert$
IEMOCAP	original	-	63.80±0.24	52.44±0.22	-44.79±0.79	-44.41±0.75	-43.68±0.63	-43.01±0.59	2.15±0.04
	Rand	data acq.	65.91±0.32	53.15±0.35	-42.38±0.93	-42.17±0.89	-41.61±0.77	-39.56±0.74	2.01±0.16
	KNN	data acq.	66.17±0.19	53.59±0.14	-39.85±0.43	-39.80±0.42	-39.02±0.38	-37.19±0.29	1.84±0.03
	Clu	data acq.	65.79±0.48	53.03±0.46	-39.04±0.73	-38.77±0.70	-38.13±0.66	-34.19±0.57	1.39±0.06
	Clu_X	data acq.	67.12±0.51	55.18±0.53	-40.23±0.69	-40.19±0.65	-39.28±0.61	-35.96±0.53	1.21±0.05
	Err	data acq.	68.19±0.26	55.44±0.27	-40.82±0.39	-40.70±0.37	-40.24±0.24	-38.97±0.19	1.75±0.04
	ours	data acq.	68.45±0.22	55.89±0.21	-33.71±0.29	-33.59±0.28	-33.01±0.21	-29.86±0.15	0.93±0.02
	Rand	data++	66.04±1.03	53.67±0.97	-41.13±1.15	-41.04±1.07	-40.55±0.89	-38.98±0.84	1.84±0.55
	KNN	data++	66.15±0.18	53.64±0.16	-39.72±0.45	-39.51±0.39	-38.79±0.35	-36.35±0.26	1.76±0.04
	Clu	data++	67.44±0.37	56.17±0.38	-36.19±0.58	-36.03±0.53	-35.28±0.41	-32.03±0.37	0.83±0.05
	Clu_X	data++	68.51±0.23	56.34±0.20	-37.71±0.48	-37.66±0.45	-36.84±0.39	-33.18±0.34	0.67±0.06
	ours	data++	**68.93±0.19**	56.41±0.16	-33.04±0.17	-32.71±0.17	-31.88±0.14	-28.93±0.11	0.59±0.03
	Rand	regular.	67.51±0.98	55.13±0.95	-40.02±1.01	-39.78±0.96	-39.11±0.82	-37.62±0.69	1.38±0.27
	KNN	regular.	68.03±0.12	55.82±0.15	-38.09±0.34	-37.95±0.31	-37.03±0.25	-35.44±0.19	1.02±0.02
	Clu	regular.	68.39±0.28	56.88±0.25	-35.41±0.47	-35.07±0.43	-34.15±0.39	-31.29±0.30	0.45±0.03
	Clu_X	regular.	68.65±0.21	56.91±0.18	-36.13±0.51	-36.10±0.49	-35.88±0.42	-32.18±0.34	0.37±0.05
	ours	regular.	68.89±0.15	**56.95±0.13**	**-32.19±0.12**	**-31.04±0.10**	**-29.57±0.09**	**-27.11±0.07**	**0.21±0.02**
	original	all data	67.15±0.13	56.13±0.17	-41.10±0.24	-40.56±0.21	-40.08±0.20	-37.11±0.14	0.88±0.02

DS	*Method*	*Strategy*	*WER*	*CER*	Δ^{-}_{max}	Δ^{-}_{avg-10}	Δ^{-}_{avg-20}	Δ^{-}_{avg-50}	$\lvert\Delta_{avg-all}\rvert$
LibriSpeech	original	-	8.05±0.05	2.80±0.04	26.11±0.98	26.02±0.95	25.57±0.89	23.11±0.76	0.29±0.06
	Rand	data acq.	7.14±0.09	2.38±0.08	17.74±0.61	17.50±0.57	17.12±0.51	16.09±0.44	0.22±0.09
	KNN	data acq.	7.03±0.04	2.32±0.06	14.95±0.47	14.73±0.41	14.19±0.35	13.81±0.32	0.13±0.04
	Clu	data acq.	6.42±0.07	2.01±0.06	12.38±0.52	12.26±0.48	12.07±0.43	11.59±0.37	0.09±0.05
	Clu_X	data acq.	6.35±0.09	2.00±0.04	13.43±0.64	13.40±0.61	12.78±0.56	12.09±0.49	0.08±0.03
	Err	data acq.	6.32±0.03	2.01±0.04	17.09±0.58	16.87±0.53	16.22±0.45	14.79±0.36	0.21±0.07
	ours	data acq.	6.31±0.04	1.99±0.04	9.51±0.36	9.38±0.29	9.02±0.25	7.87±0.16	0.07±0.03
	Rand	data++	6.89±0.15	2.25±0.14	17.44±0.57	17.28±0.53	17.07±0.42	16.01±0.34	0.17±0.10
	KNN	data++	6.41±0.07	2.12±0.04	13.19±0.32	13.11±0.26	12.64±0.21	11.38±0.11	0.12±0.06
	Clu	data++	5.95±0.08	1.92±0.09	11.72±0.38	11.48±0.32	11.09±0.24	10.65±0.20	0.08±0.05
	Clu_X	data++	5.86±0.06	1.90±0.05	12.81±0.47	12.80±0.45	12.37±0.41	11.76±0.38	0.07±0.04
	ours	data++	5.82±0.04	1.87±0.06	9.27±0.17	9.01±0.14	8.55±0.12	7.72±0.09	0.04±0.02
	Rand	regular.	6.74±0.17	2.17±0.15	17.51±0.49	17.33±0.47	16.92±0.38	15.84±0.35	0.15±0.09
	KNN	regular.	6.24±0.05	2.05±0.05	13.04±0.26	12.79±0.23	12.08±0.17	10.70±0.12	0.11±0.05
	Clu	regular.	5.80±0.07	**1.83±0.06**	10.98±0.41	10.56±0.38	10.01±0.32	9.47±0.24	0.06±0.03
	Clu_X	regular.	5.74±0.06	**1.83±0.05**	11.27±0.38	11.24±0.36	10.98±0.29	10.16±0.25	0.04±0.02
	ours	regular.	**5.71±0.07**	**1.83±0.05**	**9.12±0.11**	**8.81±0.09**	**8.14±0.08**	**7.59±0.05**	**0.03±0.02**
	original	all data	6.31±0.07	1.98±0.06	14.71±0.85	14.55±0.79	13.98±0.76	13.01±0.68	0.11±0.03

Subgroup Mitigation Analysis. We employ Global Shapley Values (GSV) to analyze how different metadata attributes contribute to model performance disparities. Figure 3.11 presents the top-15 metadata values' GSV before and after mitigation for the FSC dataset.

The random baseline fails to reduce GSV effectively, with some values actually increasing after mitigation. Clustering-based approaches show moderate success in reducing these values, demonstrating the benefit of subgroup-level mitigation. Our



(a) Regularization w/ random (b) Regularization w/ clustering (c) Regularization w/ ours

(d) Targeted data++ w/ random (e) Targeted data++ w/ clustering (f) Targeted data++ w/ ours

(g) Data acquisition w/ random (h) Data acquisition w/ clustering (i) Data acquisition w/ ours

Fig. 3.11 **Global Shapley values (GSV).** Comparison of the top-15 GSVs for the original model (orange) versus random- (shaded red, left), clustering- (shaded green, middle), and *ours* divergence-aware (shaded blue, right) weighting. **Top**: (i) in-processing regularization; **Middle**: (ii) targeted data augmentation; **Bottom**: (iii) post-processing strategy. Model: wav2vec 2.0 base (`w2v2-b`), Dataset: FSC.

approach achieves the most substantial reduction in global contributions across all scenarios. This pattern holds true for both in-processing strategies (regularization and augmentation, Figure 3.11, top and middle) and post-processing acquisition (Figure 3.11, bottom).

Full training set analysis. Table 3.12 presents results using the complete training dataset for FSC and ITALIC. This analysis focuses solely on in-processing techniques, as post-processing requires separate held-out data. The results confirm the

Table 3.12 **Mean±std over three runs on IC.** Results include original fine-tuning and in-processing mitigation strategies, namely regularization (`regular.`) and targeted data augmentation (`data++`) with $K = 2$, using *all* available training data. Best results per dataset are in **bold**, second-best underlined, and best per dataset and strategy highlighted in light blue.

DS	*Approach*	*Strategy*	*Accuracy*	*F1 Macro*	Δ^{-}_{max}	Δ^{-}_{avg-10}	Δ^{-}_{avg-20}	Δ^{-}_{avg-50}	$\lvert\Delta_{avg-all}\rvert$
FSC	original	-	93.42±0.17	93.11±0.17	-53.18±0.15	-50.89±0.09	-45.61±0.14	-40.37±0.16	0.37±0.01
	Rand.	data++	94.91±0.87	94.46±0.86	-42.62±2.94	-42.51±2.88	-41.80±2.72	-37.19±2.38	0.36±0.24
	KNN	data++	96.72±0.34	96.15±0.39	-40.01±1.59	-39.59±1.57	-38.61±1.32	-34.09±0.99	0.31±0.08
	Clu.	data++	97.85±0.37	97.59±0.65	-37.57±2.68	-37.21±2.49	-36.13±2.32	-32.75±2.07	0.24±0.11
	Clu.$_X$	data++	98.19±0.31	98.05±0.29	-38.18±2.51	-38.16±2.50	-37.45±2.44	-34.03±2.27	0.23±0.08
	ours	data++	98.46±0.11	98.42±0.17	-27.51±0.56	-27.12±0.52	-26.84±0.48	-22.15±0.43	0.21±0.08
	Rand.	regular.	96.46±0.56	96.29±0.66	-41.31±7.00	-41.31±7.00	-41.14±7.04	-40.66±7.15	0.79±0.94
	KNN	regular.	97.55±0.28	97.38±0.24	-38.29±2.34	-38.02±2.25	-36.56±2.01	-32.15±1.54	0.53±0.06
	Clu.	regular.	97.88±0.33	97.65±0.57	-36.95±8.44	-36.28±8.21	-33.69±7.24	-30.74±6.48	0.13±0.02
	Clu.$_X$	regular.	98.25±0.28	98.09±0.31	-37.86±7.15	-37.84±7.12	-36.19±7.02	-32.57±6.85	0.12±0.02
	ours	regular.	**98.47**±0.11	**98.43**±0.14	**-24.49**±0.57	**-24.49**±0.57	**-24.11**±0.51	**-22.09**±0.38	**0.11**±0.01
ITALIC	original	-	75.71±0.36	73.22±0.33	-47.54±0.79	-47.36±0.76	-46.68±0.47	-41.93±0.00	0.15±0.03
	Rand.	data++	76.06±0.29	73.36±0.77	-45.82±1.89	-45.34±1.72	-44.65±1.39	-40.82±1.10	0.13±0.09
	KNN	data++	77.15±0.21	74.03±0.24	-37.87±0.89	-37.12±0.83	-36.41±0.74	-34.04±0.67	0.12±0.04
	Clu.	data++	77.81±0.56	74.19±0.49	-36.73±2.53	-36.19±2.27	-34.15±2.02	-32.58±1.84	0.08±0.02
	Clu.$_X$	data++	77.94±0.43	74.51±0.40	-37.88±2.41	-37.84±2.38	-36.79±2.25	-33.81±2.08	0.06±0.02
	ours	data++	78.01±0.49	74.74±0.35	-30.49±1.77	-30.02±1.52	-27.48±1.47	-24.73±1.21	0.05±0.03
	Rand.	regular.	77.47±0.22	72.76±0.22	-45.11±1.41	-44.99±1.40	-44.24±1.33	-39.58±1.14	0.10±0.01
	KNN	regular.	77.96±0.19	74.12±0.23	-36.39±1.17	-36.14±1.09	-33.87±0.98	-30.05±0.91	0.07±0.02
	Clu.	regular.	78.01±0.45	74.45±0.35	-32.81±2.35	-32.73±2.32	-32.13±2.38	-28.97±2.16	0.05±0.03
	Clu.$_X$	regular.	78.05±0.51	74.63±0.52	-34.04±2.12	-34.01±2.09	-33.67±1.98	-30.18±1.82	0.04±0.02
	ours	regular.	**78.07**±0.53	**74.85**±0.30	**-30.10**±1.71	**-29.64**±1.70	**-27.31**±1.66	**-24.09**±2.19	**0.01**±0.04

superiority of our approaches over baselines in both overall and subgroup metrics. Regularization continues to outperform targeted data augmentation even with increased training data. As expected, performance improves compared to the held-out configuration shown in Table 3.10.

Sensitivity Analysis

Our sensitivity analysis examines how the choice of K (number of challenging subgroups) affects the effectiveness of our mitigation strategies. This analysis focuses on data augmentation and acquisition approaches, as regularization operates independently of K. We examine this impact through two detailed case studies: FSC for IC and LibriSpeech for ASR.

Figures 3.12 and 3.13 present our evaluation results. Each figure shows results for both post-processing acquisition (top) and in-processing targeted augmentation

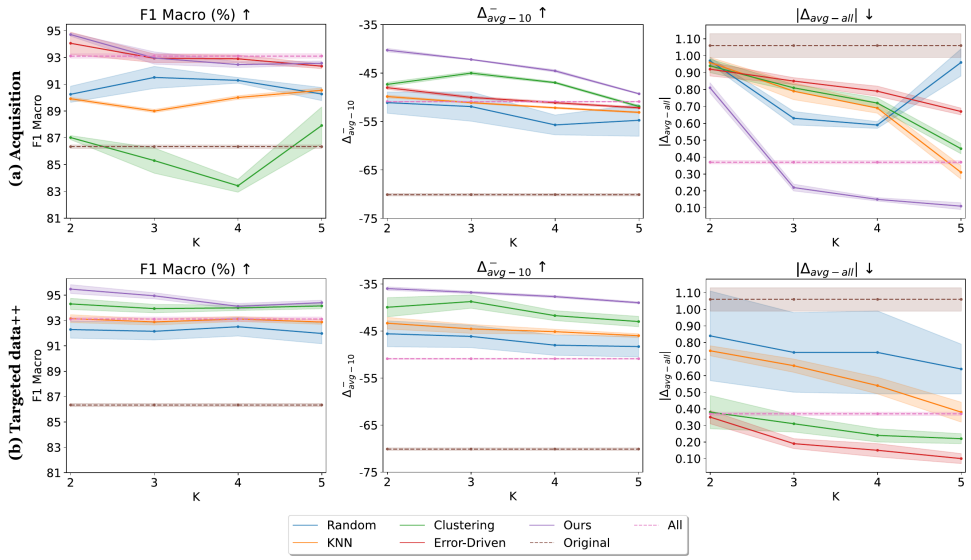


Fig. 3.12 **FSC, Sensitivity Analysis on K**. F1 Macro (left), Δ^{-}_{avg-10} (middle), and $|\Delta_{avg-all}|$ (right) for the evaluated methods in (a) post-processing acquisition (top) and (b) in-processing targeted data augmentation (bottom) setups, with K varying from 2 to 5; wav2vec 2.0 base model. Best viewed in color.

(bottom). We evaluate three key metrics: overall performance (F1 Macro/WER), subgroup divergence (Δ^{-}_{avg-10}), and system-wide divergence ($|\Delta_{avg-all}|$).

IC analysis. For the FSC dataset, our analysis reveals an inverse relationship between K and performance improvement. Lower K values lead to better overall performance by focusing mitigation efforts on the most problematic subgroups. This focused approach also reduces the average subgroup divergence (Δ^{-}_{avg-10}) more effectively. However, increasing K shows a different effect on $|\Delta_{avg-all}|$, leading to lower values. This pattern suggests that higher K values help address a broader range of subgroup behaviors, even if individual improvements are smaller.

ASR analysis. LibriSpeech results show notably different patterns from FSC. All mitigation approaches achieve better WER scores with increasing K values. This difference likely stems from the fundamental nature of the ASR task, which requires broader coverage of speech variations. Higher K values incorporate more diverse speech patterns, enhancing the generalization capabilities of the model. This improved generalization extends to subgroup performance, showing better results for top divergent subgroups. Similar to FSC, increasing K reduces $|\Delta_{avg-all}|$, indicating better system-wide fairness.

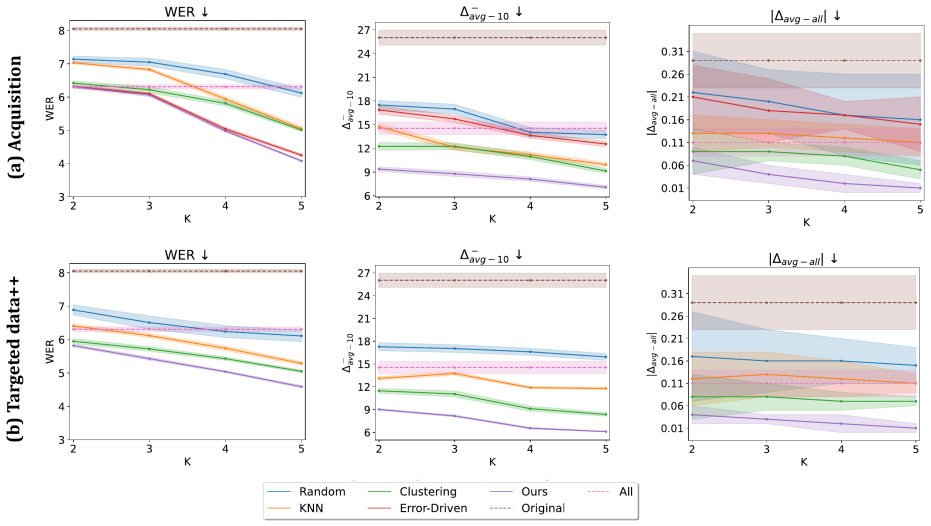


Fig. 3.13 **LibriSpeech, Sensitivity Analysis on K**. WER (left), Δ^{-}_{avg-10} (middle), and $|\Delta_{avg-all}|$ (right) for the evaluated methods in (a) the post-processing acquisition (top) and (b) in-processing targeted data augmentation (bottom) settings; Whisper base monolingual model. Best viewed in color.

Stability analysis. Our experiments reveal interesting stability patterns across multiple runs. Most baseline approaches show high variance in results, indicated by large standard deviations. However, clustering-based methods and our approach demonstrate notably more stable outcomes. This stability suggests that structured subgroup-level mitigation, whether through patterns or clusters, provides more reliable improvements.

Implications for method selection. The sensitivity analysis highlights the advantage of divergence-aware regularization over other approaches. Unlike augmentation and acquisition methods, regularization requires no K parameter tuning. This independence from K makes regularization more robust and easier to implement in practice. The analysis also emphasizes the importance of task-specific considerations when choosing K values. Simple classification tasks benefit from focused intervention (low K), while complex tasks like ASR require broader coverage (high K).

Combined Strategy Analysis

We investigate whether combining our three mitigation strategies can provide complementary benefits to improve both overall and subgroup performance. Each strategy

Table 3.13 **Joint adoption of mitigation strategies.** Results on the FSC dataset when combining the proposed strategies, including targeted data augmentation, regularization, and data acquisition.

Strategy	***Accuracy***	***F1 Macro***	Δ^{-}_{max}	Δ^{-}_{avg-10}	Δ^{-}_{avg-20}	Δ^{-}_{avg-50}	$\lvert\Delta_{avg-all}\rvert$
original	91.58±0.08	86.34±0.13	-70.09±0.26	-70.09±0.26	-65.73±0.49	-53.31±0.19	1.06±0.07
data acq.	96.55±0.08	94.71±0.12	-40.60±0.35	-40.28±0.36	-38.08±0.36	-32.72±0.28	0.81±0.03
data++	95.75±0.37	95.48±0.35	-36.12±0.39	-35.98±0.37	-34.77±0.36	-32.65±0.33	0.35±0.04
regular.	96.47±0.11	96.33±0.12	-34.49±0.45	-34.49±0.45	-34.11±0.41	-31.34±0.32	0.29±0.01
regular. & data acq.	**97.04±0.09**	96.89±0.10	-33.15±0.31	-33.12±0.29	**-32.78±0.23**	**-30.07±0.21**	0.31±0.02
data++ & data acq.	96.47±0.15	95.83±0.13	-36.14±0.36	-36.14±0.36	-35.95±0.33	-32.29±0.28	0.34±0.02
regular. & data++	96.51±0.20	96.40±0.14	-34.12±0.38	-34.10±0.38	-33.97±0.34	-30.62±0.25	0.27±0.01
regular. & data++ & data acq.	97.03±0.05	**96.91±0.04**	**-33.10±0.12**	**-33.10±0.12**	-32.82±0.09	-30.38±0.06	**0.25±0.01**
all data	93.42±0.17	93.11±0.17	-53.18±0.15	-50.89±0.09	-45.61±0.14	-40.37±0.16	0.37±0.01

addresses subgroup disparities from a different angle: data acquisition enriches the training set with underperforming samples, regularization directly penalizes subgroup divergence during training, and data augmentation increases robustness by synthetically expanding subgroup representation.

We evaluate all possible combinations on the FSC dataset: (i) regularization and data augmentation, (ii) data augmentation and data acquisition, (iii) regularization and data acquisition, and (iv) the combination of all three strategies. In each case, in-processing techniques are applied first, followed by post-processing data acquisition. The results are reported in Table 3.13.

Among pairs of strategies, the combination of regularization and data acquisition is the most effective. Regularization refines the model's performance on challenging subgroups, and acquiring additional subgroup-specific samples provides greater diversity, further improving performance beyond what either method achieves individually. Combining the two in-processing techniques slightly outperforms using them individually but is less effective than regularization + data acquisition, as both methods act on the same training data and therefore have limited complementary effect. Finally, the combination of all three strategies produces results similar to regularization and data acquisition alone. While some metrics improve further, others do not surpass the two-strategy combination. Overall, the findings indicate that while strategy combination can be beneficial, careful selection of complementary approaches is fundamental.

3.5.7 Summary and Practical Implications

Our experimental results highlight the advantages and appropriate use cases for each mitigation strategy.

Post-processing data acquisition is particularly useful when additional data can be collected to enhance an existing model. It is most suitable in scenarios with sufficient resources or budget for data collection. However, its effectiveness is limited in data-scarce or resource-constrained settings.

In-processing strategies, which operate directly on existing training data, offer greater flexibility. Among these, divergence-aware regularization consistently emerges as the most robust approach. It provides four main benefits: (i) it achieves the best overall performance across all datasets and tasks; (ii) it does not require additional data collection or augmentation; (iii) it performs stably without the need for extensive hyperparameter tuning, such as the choice of K; and (iv) it improves performance across both majority and minority subgroups.

From our results, several practical insights emerge. In-processing methods consistently outperform post-processing approaches across all metrics. Divergence-aware regularization provides the most stable and effective in-processing solution. The effectiveness of mitigation strategies depends on task complexity: simpler tasks like intent classification benefit from targeted intervention (lower K), while more complex tasks like ASR benefit from broader coverage (higher K). Combining complementary strategies, particularly regularization with data acquisition, can further enhance subgroup performance. Overall, our approaches reduce performance disparities while maintaining or improving general model accuracy.

These findings have clear practical implications for speech model development. Mitigation strategy selection should consider constraints such as data availability, computational resources, and the specific application. Our framework allows flexible deployment, enabling practitioners to optimize subgroup performance in both resource-constrained environments and scenarios where additional data collection is feasible. By explicitly addressing subgroup disparities, these strategies help ensure fairer and more reliable speech models in real-world applications.

3.6 Contrastive Learning for Bias Mitigation

Previously discussed approaches focus either on data-centric solutions (acquisition and augmentation) or regularization techniques that add specialized loss terms during training. However, these methods do not directly address how the model internally represents different subgroups, which is often the root cause of performance disparities.

Contrastive Learning (CL) has emerged as a significant advancement in representation learning [156, 157]. The core idea of CL is to learn representations that place similar samples close together and dissimilar ones further apart. We leverage this principle to guide the model in learning how to represent samples from the same subgroup close to each other [22]. The intuition is that refining the model representations at the subgroup level enables it to better capture their distinct characteristics, thus mitigating performance disparities.

3.6.1 Methodology

We propose CLUES (**C**ontrastive **L**earning for **U**nderp**E**rforming **S**ubgroups) [22], a novel framework that directly shapes the latent space of the model to better represent challenging subgroups. CLUES employs a three-level contrastive learning strategy to comprehensively address representation biases. Each level targets a different aspect of the representation space: task-level classification, subgroup characteristics, and error patterns within subgroups.

Subgroup Identification

Although our approach is agnostic to the method adopted for defining subgroups, we discuss here two possible strategies.

K-means clustering [158] applies to the latent representations of input points, extracted by the backbone model. The clustering is updated at the beginning of each training epoch to reflect the evolution of latent representations throughout training.

DivExplorer [15] leverages interpretable metadata to construct subgroups that satisfy a frequency threshold within the dataset. As discussed in previous sections, these metadata may include speaker traits (gender, age, accent), recording conditions

(duration, noise levels, speaking rate), and task-specifics characteristics (e.g., slots composing the intent). We assign each point to the most divergent subgroup it belongs to, focusing improvement efforts on underperforming subgroups.

Three-Level Contrastive Learning Framework

Our framework aims to *hint* the model, i.e., reshaping its embedding space, through three complementary objectives. First, we want embeddings that effectively separate samples according to their classification task (e.g., intent classification). Second, we seek to structure the latent space to reflect natural subgroup boundaries, ensuring distinct representations for different population segments. Third, we aim to achieve balanced representation quality across both well-performing and struggling subgroups within each classification task.

To achieve these objectives, CLUES implements three distinct contrastive learning levels: task-level, subgroup-level, and error-level contrasting. At each level, we employ the multi-similarity (MS) loss [157], which has demonstrated strong performance in speech tasks [31]. The MS loss operates by selectively contrasting pairs of samples based on their similarity relationships.

For each sample (called an *anchor* in contrastive learning terminology), we identify both *positive* samples (which should be pulled closer in the embedding space) and *negative* samples (which should be pushed apart). This push-pull dynamic shapes the embedding space by minimizing the distance to positive samples while maximizing the distance to negative ones. Each level's contribution is computed through a loss term that combines both positive and negative sample interactions:

$$\begin{aligned}\mathscr{L}_{MS} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{\alpha}\log[1+\sum_{p\in\mathscr{P}_i} e^{-\alpha(S_{ip}-\lambda)}] \\ +\frac{1}{\beta}\log[1+\sum_{n\in\mathscr{N}_i} e^{\beta(S_{in}-\lambda)}]\end{aligned} \tag{3.13}$$

where m is the batch size, $\mathscr{P}i$ and $\mathscr{N}i$ denote positive and negative sample sets, S_{ip} and S_{in} are similarities between samples, and α, β, λ control pair weighting. Figure 3.14 summarizes the action of the three contrastive loss terms on a toy example, comprised of 2 subgroups (A, B) and a binary classification task (square/triangle).

Task-level Contrastive Learning ($\mathscr{L}_t$) groups samples sharing the same class and separates samples belonging to different classes. This first loss term aims specifically at improving sample separability in the downstream task.

Subgroup-level Contrastive Learning ($\mathscr{L}_s$) guides the learning of subgroup-level representations. For this term, positive pairs are points belonging to the same subgroup, while negative pairs are points from different subgroups. This encourages an internal representation that is aware of identified subpopulations, beyond just the primary task.

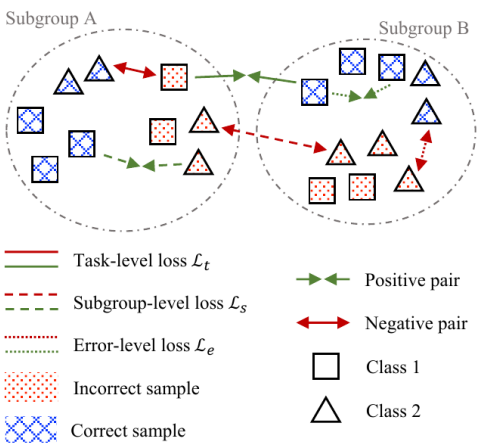


Fig. 3.14 **CLUES losses on toy example.** Illustration of how the three contrastive loss terms operate on a toy example with two subgroups (A and B) and a binary classification task (square vs. triangle).

Error-level Contrastive Learning ($\mathscr{L}_e$) considers intra-subgroup errors on the classification task. Within each subgroup, we define positive pairs as samples that have obtained the same outcome (correct/incorrect), and negative pairs as those with different outcomes. This creates a bi-partition within each subgroup: one partition containing correctly predicted samples, the other containing incorrect ones. As samples get predicted correctly, the $\mathscr{L}_e$ loss moves them toward the correctly predicted partition.

The final training objective combines these contrastive terms with standard classification loss ($\mathscr{L}_{cls}$):

$$\mathscr{L} = \mathscr{L}_{cls} + \lambda_t \mathscr{L}_t + \lambda_s \mathscr{L}_s + \lambda_e \mathscr{L}_e \tag{3.14}$$

where λ_t, λ_s, and λ_e control the relative importance of each contrastive term. This combined objective encourages the model to learn representations that are simultaneously discriminative for the task, aware of subgroup structure, and sensitive to error patterns.

3.6.2 Experimental Setup

We evaluate CLUES[9] on two intent classification datasets: English FSC and Italian ITALIC, using wav2vec 2.0 and XLS-R 128 models respectively. For subgroup identification, we set minimum frequency threshold u to 0.03 for DivExplorer, and use $K = 10$ (ITALIC) and $K = 20$ (FSC) clusters for K-means. Beyond standard accuracy and F1 metrics, we evaluate subgroup-level performance through maximum negative divergence (Δ^{-}_{max}) and representation quality through Silhouette scores. Specifically, we compute the Silhouette score with respect to the defined subgroups (S) and the partitions of correctly and incorrectly predicted samples within each subgroup ($S^{\pm}$).

We compare CLUES against our previously-discussed mitigation approaches, i.e., data acquisition (§3.4), data augmentation (§3.5), and loss regularization (§3.5). We also include two additional baselines: standard data augmentation [124], and adversarial loss for subgroup prediction[10] [108]. For both our approach and relevant baselines, we evaluate using both K-means and DivExplorer subgroup identification methods.

3.6.3 Results and Discussion

Main results. Table 3.14 presents our primary experimental findings, showing consistent patterns across both datasets.

CLUES demonstrates robust performance with both subgroup identification methods (K-Means and DivExplorer), though DivExplorer yields slightly superior results. Our approach achieves the highest overall performance metrics (accuracy and F1 score) while significantly reducing performance disparities. The improvement in fairness is particularly notable on FSC, where we reduce the maximum negative divergence from -53.2% to -17.6% for the most challenging subgroup. This reduction in performance disparity (-68% relative decrease) surpasses the improvements achieved through every other mitigation strategy.

[9] github.com/koudounasalkis/CLUES

[10] While the authors of [108] employ an extra loss for distinguishing between native and non-native speakers, we propose to discern utterances belonging to underperforming and non-underperforming subgroups.

Table 3.14 **CLUES mitigation results.** Mean±std over three runs on FSC and ITALIC. In the post-processing setup, models are trained on a reduced training set (*original - no held-out*) and use the held-out data for acquisition. In the in-processing setup, all available training data (*original*) are used before applying mitigation. We compare the *original* fine-tuning with data augmentation [124], adversarial loss [108], data acquisition [21], targeted data augmentation [20], regularization [20], and CLUES. For all metrics, higher values indicate better performance. Best results are in **bold**, second-best are underlined.

DS	*Approach*	*Subgroups*	*Accuracy*	*F1 Macro*	Δ^{-}_{max}	S	$S^{\pm}$
FSC	w2v2-b original - *no held out*	-	91.58±0.08	86.34±0.13	-70.09±0.26	0.71±0.06	0.32±0.09
	w/ data acquisition [21]	K-Means	89.77±0.88	87.02±0.15	-47.37±0.42	0.75±0.05	0.32±0.05
	w/ data acquisition [21]	DivExplorer	96.55±0.08	94.71±0.12	-40.60±0.35	0.76±0.03	0.33±0.03
	w2v2-b original - *all data*	-	93.42±0.17	93.11±0.17	-53.18±0.15	0.73±0.05	0.32±0.08
	w/ data++ [124]	-	94.91±0.87	94.46±0.86	-42.62±2.94	0.76±0.03	0.31±0.02
	w/ adversarial [108]	K-Means	98.59±0.21	98.50±0.19	-26.14±0.12	0.82±0.02	0.40±0.02
	w/ adversarial [108]	DivExplorer	98.49±0.11	98.31±0.11	-24.51±0.15	0.81±0.02	0.39±0.02
	w/ target data++ [20]	K-Means	97.85±0.37	97.59±0.65	-37.57±2.68	0.79±0.01	0.34±0.02
	w/ target data++ [20]	DivExplorer	98.46±0.11	98.42±0.17	-27.51±0.56	0.81±0.02	0.36±0.02
	w/ regularization [20]	K-Means	97.88±0.33	97.65±0.57	-36.95±8.44	0.83±0.02	0.40±0.01
	w/ regularization [20]	DivExplorer	98.47±0.11	98.43±0.14	-24.49±0.57	0.84±0.03	0.42±0.01
	w/ CLUES [22]	K-Means	98.57±0.14	98.51±0.14	-21.41±0.39	0.85±0.02	0.52±0.02
	w/ CLUES [22]	DivExplorer	**98.79±0.10**	**98.76±0.10**	**-17.58±0.43**	**0.89±0.01**	**0.53±0.01**
ITALIC	XLSR-300 original - *no held out*	-	73.79±0.32	68.08±0.37	-47.63±1.93	0.31±0.08	-0.23±0.09
	w/ data acquisition [21]	K-Means	76.31±0.51	74.02±0.51	-41.92±0.67	0.36±0.02	-0.22±0.08
	w/ data acquisition [21]	DivExplorer	77.40±0.24	72.51±0.14	-31.75±0.55	0.39±0.02	-0.20±0.02
	XLSR-300 original - *all data*	-	75.71±0.36	73.22±0.33	-47.54±0.79	0.32±0.06	-0.22±0.08
	w/ data++ [124]	-	76.06±0.29	73.36±0.77	-45.82±1.89	0.32±0.10	-0.21±0.09
	w/ adversarial [108]	K-Means	77.50±0.32	75.01±0.44	-44.12±0.65	0.45±0.10	-0.10±0.09
	w/ adversarial [108]	DivExplorer	77.20±0.64	74.84±0.53	-42.54±0.71	0.46±0.09	-0.09±0.08
	w/ target data++ [20]	K-Means	77.81±0.56	74.19±0.49	-36.73±2.53	0.47±0.08	0.03±0.05
	w/ target data++ [20]	DivExplorer	78.01±0.49	74.74±0.35	-30.49±1.77	0.48±0.09	0.04±0.04
	w/ regularization [20]	K-Means	78.01±0.45	74.45±0.35	-32.81±2.35	0.48±0.05	0.05±0.03
	w/ regularization [20]	K-DivExplorer	78.07±0.53	74.85±0.30	**-30.10±1.71**	0.49±0.06	0.05±0.07
	w/ CLUES [22]	K-Means	**80.56±0.55**	76.10±0.32	-43.01±0.89	0.51±0.03	0.21±0.03
	w/ CLUES [22]	DivExplorer	79.23±0.81	**76.72±0.20**	-40.15±0.96	**0.54±0.03**	**0.24±0.02**

We attribute these gains primarily to our subgroup-level ($\mathscr{L}_s$) and error-level ($\mathscr{L}_e$) loss terms, which effectively reduce intra-subgroup dispersion. The quality of learned representations, measured through Silhouette scores (S and $S^{\pm}$), shows substantial improvements due to the combined effect of $\mathscr{L}_s$ and $\mathscr{L}_e$ terms.

Ablation analysis. Table 3.15 details our ablation study examining the contribution of each loss term. Each individual loss term improves accuracy and F1 scores compared to the baseline model. The combination of all terms yields the strongest overall performance.

Table 3.15 **Ablation study.** Three contrastive loss components: $\mathscr{L}_t$, $\mathscr{L}_s$, and $\mathscr{L}_e$. Subgroups are extracted using DivExplorer. Best results are in **bold**.

DS	*Approach*	*F1 Macro*	Δ^-_{max}	S	$S^{\pm}$
FSC	w2v2-b	93.11	-53.18	0.74	0.312
	w/ $\mathscr{L}_t$	98.16	-48.71	0.76	0.31
	w/ $\mathscr{L}_s$	98.43	-26.55	0.85	0.42
	w/ $\mathscr{L}_t + \mathscr{L}_s$	98.43	-33.12	0.81	0.37
	w/ $\mathscr{L}_s + \mathscr{L}_e$	98.45	-19.11	0.86	0.50
	w/ $\mathscr{L}_s + \mathscr{L}^*_e$	98.11	-20.01	0.87	0.45
	w/ CLUES	**98.76**	**-17.58**	**0.89**	**0.53**
ITALIC	XLSR-300	73.22	-47.54	0.32	-0.22
	w/ $\mathscr{L}_t$	76.08	-49.54	0.35	-0.22
	w/ $\mathscr{L}_s$	76.33	-45.39	0.45	-0.20
	w/ $\mathscr{L}_t + \mathscr{L}_s$	76.28	-47.44	0.44	-0.20
	w/ $\mathscr{L}_s + \mathscr{L}_e$	76.62	-43.29	0.49	0.22
	w/ $\mathscr{L}_s + \mathscr{L}^*_e$	76.32	-43.33	0.50	0.12
	w/ CLUES	**76.72**	**-40.15**	**0.54**	**0.24**

The introduction of $\mathscr{L}_s$ produces the most significant reduction in divergence (Δ^-_{max}) by creating more separable latent representations for different subgroups. Adding $\mathscr{L}_e$ further reduces divergence while improving overall performance metrics.

We also explore an alternative error-level loss ($\mathscr{L}^*_e$) that merges correct and incorrect predictions instead of separating them. While this alternative still shows positive results, the original $\mathscr{L}_e$ proves more effective. As anticipated, the error-level loss produces the largest improvements in the $S^{\pm}$ metric.

Qualitative assessment. Figure 3.15 provides a t-SNE [159] visualization comparing original wav2vec 2.0 embeddings with CLUES-enhanced representations on FSC. The visualization focuses on the five most frequent intents and their three most frequent subgroups.

While both approaches cluster samples by intent, the original embeddings show less cohesive grouping. The baseline model struggles to place correctly and incorrectly predicted samples from the same intent near each other. Additionally, the original space fails to maintain proximity between samples from the same subgroup.

CLUES, by contrast, creates more consistent groupings across all organizational levels: intents, subgroups, and prediction correctness. The "*increase heat kitchen*" intent (yellow) particularly demonstrates this improvement. In the original space,

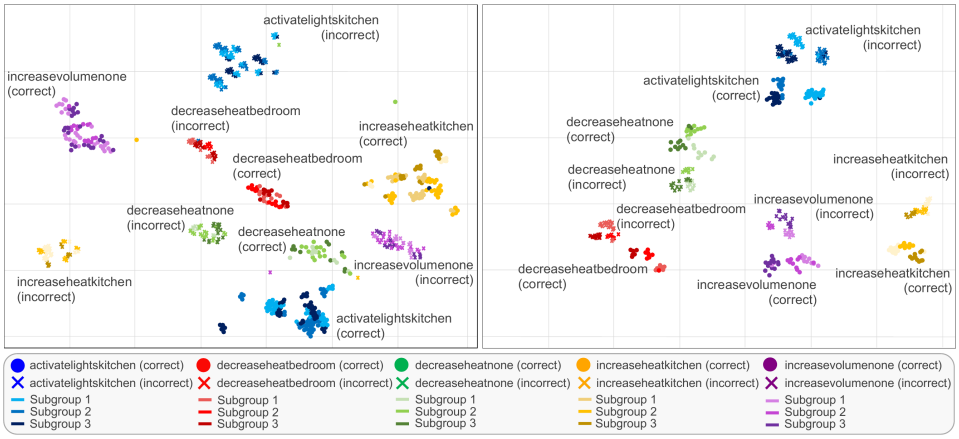


Fig. 3.15 **FSC dataset, t-SNE visualization.** Original model (left) and CLUES (right), showing the five most frequent intents (distinct colors) and the three most frequent subgroups (different shades of each color). Correct predictions are shown as circles, incorrect ones as crosses. Best viewed in color.

these samples appear scattered, with correct and incorrect predictions widely separated and poor subgroup distinction. The CLUES representation organizes these samples into a coherent three-level structure: (i) all intent samples cluster together (yellow cluster); (ii) correct and incorrect predictions form distinct partitions within the cluster, and (iii) individual subgroups create cohesive sub-clusters within each partition.

3.6.4 Summary and Practical Implications

The results show that CLUES improves both model performance and subgroup fairness across all evaluated tasks. CLUES helps the model learn more balanced and robust representations by aligning feature spaces across classes and subgroups. Both quantitative metrics, such as the Silhouette score, and qualitative analyses, such as the t-SNE visualizations, confirm this improvement. This validates that contrastive learning can be an effective direction for promoting fairness in speech models.

The ablation study supports the contribution of each loss term. When combined, the three contrastive losses provide the most consistent and stable performance gains. This confirms that CLUES effectively addresses subgroup disparities without reducing overall task accuracy.

From a practical perspective, CLUES can be easily applied to existing training pipelines. It does not require collecting new data or using external supervision. It operates only on available training samples and subgroup information. It is designed as a modular loss function that can be integrated into any classification model. In addition, it enables transparent subgroup-level analysis and helps practitioners identify and mitigate performance gaps during training.

In this work, we evaluated CLUES on intent classification. However, the same approach can be applied to other classification-based tasks. Future work could extend CLUES to ASR and other supervised speech tasks. Another direction could also be integrating automatic subgroup discovery to support scenarios where subgroup labels are not available.

3.7 Privacy-Preserving Approaches to Bias Mitigation

All the approaches to identifying and addressing these performance disparities described so far, while effective, rely heavily on demographic attributes like age, gender, and accent. However, collecting such sensitive information during model deployment raises significant privacy concerns and may be impractical or legally restricted in many scenarios. This section presents a privacy-preserving approach that enables effective bias detection and mitigation without requiring demographic metadata at inference time [23, 24].

3.7.1 Motivation

Confidence Models (CMs) have long been central to evaluating speech model reliability, producing confidence scores for both individual words and entire utterances [138–141, 160, 161]. Recent studies have demonstrated that these confidence scores can be exploited to identify problematic cohorts without relying on sensitive demographic information [3].

Building on this direction, the work in [4] represents, to the best of our knowledge, the first attempt to disentangle speaker-related and acoustic factors for problematic subgroup identification. Their method applies clustering on utterance-level embeddings extracted from a speaker identification model trained on a public dataset. While

promising, this approach presents two main limitations. First, the embedding representations are not interpretable, making the resulting clusters difficult to analyze and act upon. Second, the clustering process itself is sensitive to noise and embedding variability, which can reduce robustness and stability.

In contrast, our work aims to achieve subgroup discovery that is interpretable, stable, and does not rely on explicit demographic attributes at deployment time. We thus address these issues through a novel framework that combines the predictive power of confidence models with interpretable subgroup identification [23, 24]. The key insight is that by training a CM to recognize patterns associated with challenging subgroups during training, we can later identify similar patterns without accessing sensitive attributes. This approach enables both effective bias mitigation and privacy preservation, making it particularly suitable for real-world deployment scenarios.

3.7.2 Methodology

Given a speech model *M* trained for a specific task (e.g., IC or ASR), our goal is to improve its performance on challenging subgroups while preserving privacy. Specifically, we aim to identify potentially problematic or challenging utterances during deployment without requiring sensitive demographic information. This requires both detecting challenging samples early and providing interpretable explanations of why they might be problematic.

Our solution introduces a Challenging Subgroup Identification (CSI) model that leverages confidence scores to identify challenging cases without accessing sensitive attributes. The complete pipeline, illustrated in Figure 3.16, consists of five main steps.

1) *Confidence Model Training.* We first train a confidence model to capture patterns associated with model errors. This CM learns to predict whether utterances will be challenging for model *M* based on non-sensitive features like acoustic properties and output probabilities.

2) *Problematic Subgroup Extraction.* During training, we use DivExplorer to identify subgroups where *M* underperforms. This process uses all available metadata, including demographic information, to create interpretable descriptions of challenging populations. These descriptions serve as training targets for the CSI model.

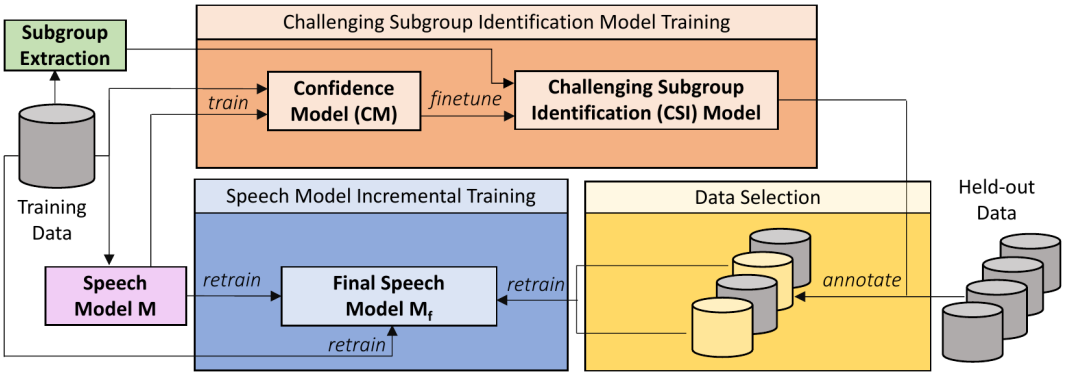


Fig. 3.16 **Overview of the proposed pipeline.** We fine-tune a Confidence Model (`CSI`) to predict the challenging subgroup to which each utterance belongs [23]. Utterances from the held-out set identified as challenging by `CSI` are then added to the original training data, enabling incremental retraining of the speech model [24].

3) *CSI Model Training.* We fine-tune the CM to predict whether utterances belong to the identified challenging subgroups. Importantly, while the model learns from demographically-aware subgroups during training, it makes predictions using only privacy-preserving features.

4) *Privacy-Preserving Data Selection.* Using the trained CSI model, we identify potentially challenging utterances in new, unlabeled data. This selection process requires no demographic information, relying instead on the patterns learned by the CSI model.

5) *Targeted Model Improvement.* Finally, we retrain model M using the original training data augmented with the selected challenging samples. This targeted update helps improve performance on difficult subgroups while maintaining privacy.

The key innovation of this pipeline is its ability to leverage demographic insights during training while operating in a completely privacy-preserving manner during deployment.

Confidence Model Training

The first step in our pipeline involves training a confidence model to recognize patterns associated with model errors. For a given speech model M, we create a transformed dataset Z from the original utterances X by extracting three types of features, including (i) *uncertainty indicators*, i.e., N-best list lengths and output probability

distributions, (ii) acoustic representations, i.e., acoustic embeddings representing the (last or average of the) hidden states of the model, and (iii) non-sensitive speech features, i.e., word counts, speaking rate, pause patterns and durations.

For each utterance, we assign binary labels indicating model performance: 1 if the prediction by the model M is correct, 0 otherwise. For ASR specifically, "correct" means achieving a WER of 0.0.

The CM is trained using standard supervised learning procedures. The dataset Z is split into train, validation, and test sets. Model parameters are optimized on the training and validation sets, while the test set provides final performance evaluation. This training process equips the CM with the ability to recognize error patterns, the domain-specific knowledge about the task, and an understanding of the behavior of model M. Importantly, while demographic information may be used during training to identify challenging cases via DivExplorer, the CM learns to make its predictions using only non-sensitive features.

Challenging Subgroup Identification

The second step involves identifying which subgroups of the population present the greatest challenges for model M. We employ DivExplorer to analyze model performance across different population segments. During this training phase, we deliberately use all available metadata, including demographic information, speech characteristics, and task-specific features.

DivExplorer allows us to identify subgroups where model performance significantly diverges from the average. We select the top-K most challenging subgroups based on negative divergence values, applying redundancy pruning to ensure diverse and non-overlapping subgroup definitions (§3.4). While this step uses demographic information, we note that it does so only during training to create interpretable descriptions of challenging cases. These descriptions will guide the CSI model in learning to recognize similar patterns without accessing sensitive attributes.

CSI Model Training

The third step creates our privacy-preserving detector by fine-tuning the previously trained CM to recognize challenging subgroups. Using the transformed dataset Z, we

assign each utterance one of two labels, either a specific problematic subgroup ID, or a "Non-problematic" label (0). For utterances belonging to multiple challenging subgroups, we assign the ID of the subgroup showing the highest performance divergence.

The CSI model learns to make these predictions using only the privacy-preserving features it has been trained with, i.e., the speech metadata (duration, rate, pauses), the model acoustic embeddings (hidden states), and the model output probabilities. This training process enables the CSI model to recognize patterns associated with challenging subgroups without requiring access to the demographic information used to initially identify those subgroups. The result is a binary classifier that can flag potentially challenging or problematic utterances during deployment while maintaining privacy.

Privacy-Preserving Data Selection and Model Update

The final steps of our pipeline focus on using the CSI model to improve the performance of model M across all subgroups while maintaining privacy.

Instead of indiscriminate data acquisition, we implement a targeted approach similar to the one presented in §3.4: (i) we start with a held-out set of unlabeled utterances; (ii) we use the CSI model to identify potentially challenging samples; and (iii) we select these samples for annotation and inclusion. At this point, we combine the original training data with the newly selected samples, and we fine-tune model M on this augmented dataset.

This targeted approach efficiently addresses performance disparities while still maintaining privacy by avoiding demographic data collection, and enables continuous model improvement in production settings. The entire process operates without requiring sensitive attributes during deployment, making it suitable for real-world applications where privacy is fundamental.

3.7.3 Experimental Setup

We evaluate our approach[11] across two datasets and two tasks: FSC for IC and LibriSpeech for ASR. We employ wav2vec 2.0 base (FSC) and Whisper base monolingual (LibriSpeech).

Confidence model. The CM architecture features two hidden layers with GELU activation function, dropout, and normalization layers. We initialize the layers through the Kaiming normal initialization technique. We train the model with Cross-Entropy (CE) loss for the IC task, whereas for ASR we implement a Mean Squared Error (MSE) component using WER as an extra target. As a result, the objective function includes a weighted combination of CE and MSE losses, described by: $\mathscr{L}_{tot} = \alpha\mathscr{L}_{CE} + (1-\alpha)\mathscr{L}_{MSE}$ where α is set to 0.6.

Baselines. We compare our mitigation strategy against the same baselines introduced in Section §3.5. Specifically, we benchmark our acquisition approach with those guided by (i) random selection, (ii) KNN-based selection, (iii) clustering-based selection, and (iv) CM-based error prediction. In addition to these baselines, we further utilize two oracle approaches that serve as theoretical upper bounds by using information unavailable in real deployment. The former is a *Supervised Oracle* (`S-Oracle`), which is based on the methodology from [155]. It directly identifies erroneous predictions using ground truth labels, thus serves as performance ceiling for our CM-error-based selection method. However, it requires access to true labels, which are typically unavailable during deployment. The latter is a *Metadata-based Oracle* (`M-Oracle`), which follows the approach described in Section §3.4. It uses full demographic and sensitive metadata for selection, thus representing an ideal subgroup-aware data acquisition scenario and providing an upper bound for our CSI approach while highlighting privacy trade-offs.
These oracles help quantify the performance gap between privacy-preserving methods and approaches with unlimited access to sensitive information.

[11] github.com/koudounasalkis/CMs-for-Problematic-Subgroups
github.com/koudounasalkis/CSI-for-Bias-Mitigation

Table 3.16 **CM and CSI results on FSC and LibriSpeech.** Performance of the Confidence Model (Step 1) and the `CSI` model for challenging subgroup identification (Step 3) on the FSC and LibriSpeech (LS) datasets. Best results are shown in **bold**.

	CM Performance		Challenging Subgroups Identification							
	AUC	Accuracy	Approach	K = 2		K = 3		K = 4		K = 5	
				ERR ↓	F1 ↑	ERR ↓	F1 ↑	ERR ↓	F1 ↑	ERR ↓	F1 ↑
FSC	0.74	88.85%	Random (uniform)	67%	22%	75%	14%	80%	11%	83%	7%
			Random (majority)	10%	32%	13%	23%	16%	18%	21%	14%
			KNN	6%	78%	8%	66%	11%	60%	12%	62%
			CSI w/out CM pretrain	10%	32%	12%	27%	15%	25%	16%	26%
			CSI	**4%**	**88%**	**6%**	**77%**	**8%**	**75%**	**8%**	**77%**
LS	0.73	74.54%	Random (uniform)	67%	32%	75%	23%	79%	17%	83%	14%
			Random (majority)	56%	20%	60%	14%	62%	13%	67%	10%
			KNN	18%	68%	31%	50%	32%	50%	43%	37%
			CSI w/out CM pretrain	32%	53%	39%	47%	41%	40%	42%	30%
			CSI	**16%**	**83%**	**24%**	**58%**	**29%**	**54%**	**31%**	**50%**

3.7.4 Results and Discussion

Subgroup Identification Performance

Table 3.16 reports the results of CM error identification (Step 1) and challenging subgroup identification (Step 3) on the FSC and LibriSpeech datasets. We evaluate our approach against three baseline strategies. The first baseline randomly assigns classes to utterances, the second assigns all samples to the majority class, and the third uses a KNN-based assignment. Our method clearly outperforms all baselines, highlighting the effectiveness of our approach for accurately identifying challenging subgroups.

The CM achieves strong predictive performance, with AUC scores of 0.74 and 0.73 for FSC and LibriSpeech respectively. While these AUC values might not appear exceptionally high, they prove more than sufficient for our purposes. Using these pre-trained CMs as a foundation for fine-tuning on the subgroup identification task leads to substantial improvements in both Error Rate and F1 macro scores compared to training the CSI model from scratch.

Examining specific results, the FSC dataset with $K = 2$ provides a clear illustration. The CSI model accurately identifies challenging subgroups defined by particular combinations of speech and task-related metadata. For example, it detects utterances with low word counts targeting "*heat*" objects and short-duration utterances involving "*decrease*" actions. Initializing the CSI model from CM pretraining significantly

improves performance compared to training from scratch: F1 scores rise from 32% to 88%, and the equal error rate (EER) decreases from 10% to 4%. This corresponds to more than a 50% gain in subgroup identification accuracy. Similar patterns emerge in the LibriSpeech dataset, where we observed a 50% EER reduction and a huge increase in F1 macro score from 53% to 83%. This consistent improvement across different datasets suggests the robustness of our approach.

Importantly, these performance gains are stable when varying K from 2 to 5 subgroups. In all cases, the CSI model initialized from our pre-trained CM consistently outperforms models trained from scratch. This demonstrates that the CM effectively captures information about error patterns that can be leveraged for subgroup identification. These findings highlight the critical role of the confidence pre-training strategy in building accurate and privacy-preserving subgroup identification systems.

Optimal feature configuration. In developing our classification model, we conduct extensive experiments to determine the optimal feature configuration. We explore various combinations of features, including model logits, output probabilities, n-best sequence lengths (particularly relevant for LibriSpeech), speech metadata, and hidden state representations. While the best performance is achieved using all features comprehensively, our analysis reveals interesting task-specific patterns. For FSC, output probabilities emerge as the dominant factor driving performance improvements. In contrast, for LibriSpeech, leveraging the hidden states contributes most significantly to achieving the best performance.

Bias Mitigation Results

Table 3.17 presents the evaluation of our privacy-preserving data selection approach on the FSC dataset. The results demonstrate the ability of our method to significantly reduce performance disparities while maintaining or improving overall model performance. When examining the impact on challenging subgroups, we observe a reduction in the Intent Error Rate (IER) of approximately 50% for $K = 2$ and more than 60% for $K = 5$ compared to the original fine-tuned model. This improvement in subgroup performance is accompanied by substantial gains in overall metrics, with a 39% reduction in IER and nearly 10% improvement in F1 macro scores. These results consistently outperform all baseline approaches across every value of K we evaluated.

Table 3.17 **CSI mitigation results on FSC, wav2vec 2.0 base.** Mean±std over three runs. `K` denotes the number of challenging subgroups considered, and `N` the number of selected samples. We compare the `Original` fine-tuning, baseline methods, our `CSI`, and the two oracles (`M-Oracle` and `S-Oracle`). Best results for each K are highlighted in light blue, and best results with oracles are in **bold**.

K	N	Approach	IER (%) ↓	F1 Macro (%) ↑	IER top-K (%) ↓
-	18506	Original	8.42±0.08	86.34±0.13	67.63±0.08 (K = 2)
2	+223	Random	9.19±0.03	88.48±0.05	65.90±0.22
		KNN	7.93±0.07	89.92±0.10	59.90±0.23
		Clustering [3]	7.06±0.07	91.82±0.15	47.35±0.42
		CM	6.87±0.04	93.93±0.05	52.24±0.35
		CSI (*ours*)	5.17±0.03	**94.87±0.03**	34.04±0.21
		S-Oracle [155]	5.29±0.02	94.06±0.04	47.47±0.39
		M-Oracle [21]	**4.46±0.08**	94.81±0.09	**32.95±0.36**
-	+4606	All data	6.58±0.17	93.11±0.17	55.11±0.24 (K = 2)
3	+361	Random	9.41±0.05	88.15±0.09	49.44±0.38
		KNN	8.25±0.09	89.12±0.14	39.30±0.36
		Clustering [3]	7.19±0.06	91.06±0.09	37.15±0.39
		CM	6.15±0.05	92.30±0.07	38.80±0.43
		CSI (*ours*)	5.25±0.04	94.21±0.07	23.17±0.23
		S-Oracle [155]	5.60±0.04	93.43±0.04	51.17±0.35
		M-Oracle [21]	**5.12±0.04**	**94.41±0.06**	**22.89±0.12**
4	+397	Random	9.45±0.11	88.09±0.10	36.44±0.27
		KNN	8.29±0.02	89.51±0.07	25.50±0.29
		Clustering [3]	7.42±0.07	90.89±0.08	36.08±0.31
		CM	6.59±0.04	91.75±0.05	38.19±0.25
		CSI (*ours*)	5.31±0.03	94.19±0.05	19.89±0.21
		S-Oracle [155]	5.84±0.06	93.44±0.06	46.40±0.33
		M-Oracle [21]	**5.19±0.06**	**94.25±0.07**	**18.72±0.17**
5	+467	Random	9.58±0.10	88.04±0.10	34.80±0.39
		KNN	8.31±0.03	89.50±0.06	21.24±0.23
		Clustering [3]	7.68±0.06	90.61±0.05	29.75±0.27
		CM	6.70±0.05	91.69±0.03	25.34±0.23
		CSI (*ours*)	5.39±0.06	94.05±0.04	14.55±0.08
		S-Oracle [155]	5.85±0.06	**94.76±0.03**	46.94±0.25
		M-Oracle [21]	**5.28±0.04**	94.08±0.06	**14.01±0.11**
-	+4606	All data	6.58±0.17	93.11±0.17	39.78±0.12 (K = 5)

A critical aspect of our evaluation involves comparing against two oracle approaches that have access to information unavailable in real-world deployments. The metadata-based oracle (M-Oracle) uses complete demographic information, while the supervised oracle (S-Oracle) leverages ground truth labels. Our method proves remarkably competitive with these theoretical upper bounds. Particularly when compared to M-Oracle, our approach achieves comparable results despite operating without access to sensitive demographic information. This is especially significant given that many of the most challenging subgroups involve demographic attributes. For instance, one of the top-2 challenging groups in FSC consists of male speakers aged 41-65 with

high speaking rates, yet our method successfully identifies and addresses such cases without directly accessing this demographic data.

When compared to S-Oracle, which uses ground truth labels to identify problematic samples, our approach shows slightly better performance, particularly in reducing IER for top-K subgroups. We attribute this advantage to our method's ability to recognize and target specific population subgroups, rather than just focusing on individual errors. While S-Oracle can identify which samples will be misclassified, it lacks the broader perspective of systematic disparities across population segments that our approach captures.

Table 3.18 **CSI mitigation results on LibriSpeech, whisper base.** Mean±std over three runs. Best results for each K are highlighted in light blue, and best results with oracles are in **bold**.

K	N	Approach	WER ↓	WER top-K ↓
-	83211	Original	8.05±0.05	25.91±0.98 (K = 2)
2	+6912	Random	7.96±0.29	25.02±0.44
		KNN	7.80±0.04	18.44±0.32
		Clustering	7.33±0.08	14.05±0.38
		CM	7.70±0.09	14.86±0.27
		CSI (*ours*)	7.25±0.06	**12.33±0.15**
		S-Oracle	7.28±0.09	24.17±0.29
		M-Oracle	**7.22±0.06**	12.51 ±0.09
-	+20803	All data	6.31±0.07	17.46±0.87 (K = 2)
3	+8120	Random	7.71±0.31	22.15±0.41
		KNN	7.55±0.05	16.29±0.28
		Clustering	7.08±0.10	13.09±0.31
		CM	7.49±0.07	13.01±0.23
		CSI (*ours*)	6.81±0.08	10.97±0.17
		S-Oracle	6.87±0.07	21.86±0.32
		M-Oracle	**6.80±0.05**	**10.94±0.11**
4	+9958	Random	7.40±0.24	20.43±0.33
		KNN	7.33±0.04	14.84±0.19
		Clustering	6.81±0.08	12.55±0.24
		CM	7.21±0.05	12.56±0.18
		CSI (*ours*)	6.48±0.07	10.16±0.15
		S-Oracle	6.47±0.09	19.74±0.29
		M-Oracle	**6.43±0.05**	**10.15±0.09**
5	+12026	Random	7.14±0.09	17.52±0.31
		KNN	7.03±0.04	12.77±0.16
		Clustering	6.42±0.07	11.19±0.26
		CM	6.81±0.05	11.04±0.19
		CSI (*ours*)	6.32±0.04	9.33±0.13
		S-Oracle	6.34±0.05	15.01±0.26
		M-Oracle	**6.31±0.04**	**9.32±0.08**
-	+20803	All data	6.31±0.07	12.24±0.79 (K = 5)

Similar patterns emerge in our LibriSpeech evaluation, as shown in Table 3.18. Our method achieves the lowest overall WER of 6.32 and reduces the WER for the most challenging subgroups to 9.33 (with $K = 5$). These results either match or exceed the performance of both oracle approaches, demonstrating that privacy-preserving methods can be just as effective as those using sensitive information.

The LibriSpeech results also highlight an important efficiency advantage of our approach. Using only 60% of the available held-out data, we achieve performance comparable to utilizing the complete dataset. More significantly, our targeted selection strategy proves more effective than indiscriminate data addition: we achieve a top-K WER of 9.33, substantially

outperforming the 12.24 WER obtained when simply adding all available data. This demonstrates that intelligent, privacy-preserving data selection can be more valuable than raw data quantity. While our ASR results may not represent absolute state-of-the-art performance, they clearly demonstrate how privacy-preserving, targeted data selection can effectively reduce bias while improving overall performance. Using Whisper base as our foundation, we show that thoughtful data selection can be as important as model architecture for achieving fair and robust performance across diverse speaker populations.

3.7.5 Summary and Practical Implications

Our evaluation shows that privacy-preserving data selection can effectively reduce subgroup performance disparities without harming overall model accuracy. The improvements are consistent across multiple tasks, including IC and ASR, as well as different model architectures, such as wav2vec 2.0 and Whisper. Importantly, our approach achieves performance comparable to or better than oracle methods that rely on sensitive demographic information or exact error labels. This demonstrates that it is not necessary to access demographic data at deployment to mitigate bias effectively. By leveraging confidence scores and model-derived signals, the method identifies systematically underperforming subgroups and targets them for mitigation, all while preserving user privacy.

From a practical standpoint, this approach is particularly useful in real-world scenarios where collecting or using demographic information is legally or ethically restricted. It also reduces the need for extensive manual labeling, allowing practitioners to improve fairness efficiently with limited additional data. Finally, the method can be adapted to different tasks and model architectures, making it a flexible tool for promoting equitable performance in diverse speech processing applications.

3.8 Conclusions

This chapter has presented a comprehensive framework for analyzing and mitigating subgroup performance disparities in speech models. We introduced five complementary approaches that address different aspects of the fairness challenge, from discovery to mitigation.

Our divergence analysis framework (§3.3) provides interpretable insights into where and how models fail, enabling systematic identification of challenging subgroups without relying on predetermined categories. The analysis revealed that performance disparities often emerge from complex interactions between multiple attributes, such as when gender intersects with speaking rate or when task-specific features combine with acoustic conditions.

For mitigation, we first introduced a post-processing strategy based on divergence-aware data acquisition (§3.4). This approach strategically selects new training samples from challenging subgroups, achieving better results with less additional data compared to indiscriminate data collection. The effectiveness of this targeted acquisition demonstrates that intelligent data selection can be more valuable than raw data quantity.

We then developed two complementary in-processing techniques (§3.5). Targeted data augmentation enriches the training data for challenging cases, while divergence-aware regularization adjusts the training objective of the model to focus on underperforming subgroups. Our experiments showed that these in-processing methods consistently outperform post-processing approaches, with regularization emerging as particularly effective due to its parameter efficiency and stability.

The contrastive learning approach, CLUES, introduced in Section §3.6, takes a different angle by directly improving model representations for underperforming subgroups. Through its three-level learning strategy, it achieves more equitable performance by reshaping the model latent space to better capture both task requirements and population characteristics.

Finally, our privacy-preserving framework (§3.7) demonstrates that effective bias detection and mitigation is possible even without accessing sensitive demographic information during deployment. By leveraging confidence models to identify challenging cases, we achieve overall- and subgroup-level performance improvements comparable to methods that use demographic data, while maintaining user privacy.

Our experimental results across multiple tasks (IC, SER, ASR), languages (English, Italian), and architectures (wav2vec 2.0, XLS-R, HuBERT, Whisper) demonstrate the generality and effectiveness of these approaches.
Each method shows distinct advantages: (i) in-processing techniques offer the most systematic improvements; (ii) targeted data acquisition provides cost-effective mitigation for existing models; (iii) contrastive learning creates more equitable

internal representations; and (iv) privacy-preserving methods enable real-world deployment while protecting sensitive information.

Together, these contributions provide a practical pathway toward more equitable speech technology, offering solutions adaptable to different scenarios and constraints. Whether working with existing models, limited data, or privacy requirements, our framework provides effective tools for improving fairness while maintaining or enhancing overall performance.

These advances have significant implications for deploying speech technology in sensitive domains like healthcare, legal transcription, and emergency services. By enabling bias mitigation without compromising privacy or performance, our framework addresses critical barriers to responsible AI deployment.

3.8.1 Future Research Directions

Our work suggests several promising directions for future research.

In subgroup discovery, incorporating temporal dynamics into divergence analysis could help understand how performance disparities evolve during model training and deployment.

For mitigation strategies, exploring adaptive techniques that automatically balance different approaches based on subgroup characteristics could lead to more efficient solutions. Investigating the relationship between model architecture and bias susceptibility could inform the design of inherently fairer models. Research into few-shot adaptation for challenging subgroups might enable rapid performance improvements without extensive data collection.

In privacy preservation, developing techniques that can transfer knowledge about challenging subgroups across different tasks while maintaining privacy could enhance the practical utility of these methods. Exploring the connection between confidence modeling and representational bias could lead to more sophisticated privacy-preserving detection approaches.

These advances would contribute to developing speech technology that is not only powerful but also fair, privacy-preserving, and suitable for real-world deployment across diverse populations.

Chapter 4

Frameworks for Evaluating Speech Foundation Models

4.1 Introduction: The Need for Multi-Dimensional Evaluation

Foundation models have set new performance standards across the field of speech processing. Their success is typically measured by aggregate metrics like WER or overall accuracy. These metrics are useful for tracking general progress. However, these single scores provide an incomplete and often misleading picture of model behavior. They treat all errors as equal, ignoring the vast differences in their real-world impact. For instance, a minor phonetic mistake and a critical semantic inversion are penalized similarly by WER. This approach does not capture the critical aspects of performance needed for safe and reliable deployment.

As these models move from research labs into real-world applications, these evaluation gaps become more significant. A model with a low WER might still generate fluent but entirely fabricated text, a phenomenon known as *hallucination*. A model with high classification accuracy may not have mechanisms to forget user data upon request, failing to comply with privacy regulations. And a model pre-trained exclusively on speech may not generalize well to other types of audio, limiting its utility as a true "*foundation*" model. This growing gap between benchmark scores and real-world reliability highlights a fundamental problem in the field. We cannot build

truly robust, responsible, and trustworthy systems if we cannot properly measure these qualities.

This chapter addresses the critical need for more comprehensive and nuanced evaluation methodologies. We argue that the next stage of progress in speech technology requires moving beyond simple, monolithic accuracy metrics. We need specialized, multi-dimensional benchmarks that can assess models along distinct axes of performance. To this end, we introduce three novel benchmarking frameworks developed as part of this thesis. Each framework is designed to provide a deeper, more granular evaluation of speech foundation models, targeting a critical aspect of their behavior.

First, we address the problem of model trustworthiness by focusing on the detection and characterization of hallucinations. We present SHALLOW [25], the first benchmark to systematically categorize and quantify hallucinations in automatic speech recognition. It moves beyond simple error counting to provide a fine-grained analysis of model outputs. SHALLOW assesses errors across four complementary dimensions: lexical, phonetic, morphological, and semantic. This allows us to distinguish between minor inaccuracies and severe, meaning-altering fabrications.

Second, we tackle the challenge of model responsibility and data privacy, which are central to building trustworthy AI. We introduce UnSLU-BENCH [26], the first benchmark for evaluating machine unlearning in spoken language understanding. This framework provides a standardized way to measure the ability of a model to effectively forget specific speaker data. It supports the "*right to be forgotten,*" a key principle in modern data protection laws. UnSLU-BENCH also proposes a novel metric to holistically evaluate unlearning on its efficacy, utility, and computational efficiency.

Third, we examine the core quality of foundation models: the generalizability of their learned representations. We present ARCH [27], a unified and extensible benchmark for evaluating audio representation learning across diverse domains. Current benchmarks often focus narrowly on speech-related tasks. ARCH, in contrast, assesses model performance on speech, music, and environmental sound tasks. This provides a holistic measure of the ability of a model to serve as a foundation for a wide range of audio-centric applications.

Together, these three frameworks form a comprehensive suite for evaluating modern speech and audio models. They collectively push the field beyond a narrow focus on transcription accuracy and towards a more complete understanding of model

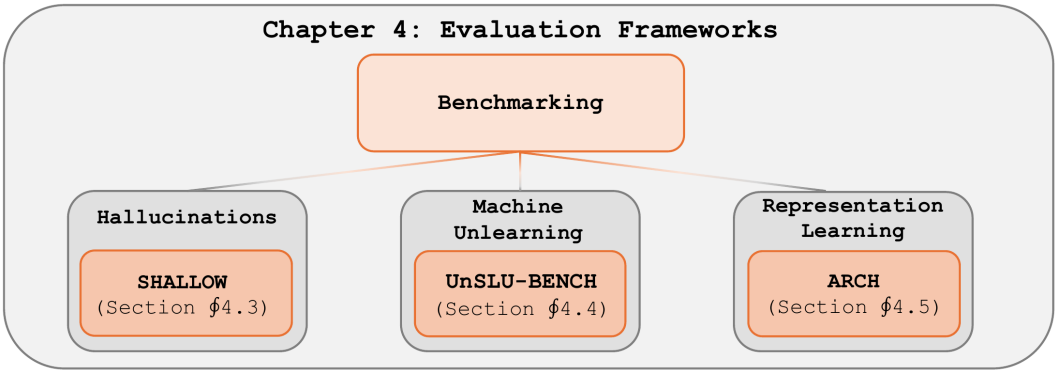


Fig. 4.1 **Chapter 4 Overview.** Graphical taxonomy of Chapter 4 topics.

behavior. They provide the specific tools needed to measure robustness against hallucinations, compliance with privacy principles, and the true generalization of learned knowledge. By enabling a deeper assessment of model strengths and weaknesses, these benchmarks allow for the development of the next generation of speech AI. This next generation of models will be safer, more reliable, and better aligned with the complex demands of real-world applications.

The remainder of this chapter is organized as follows. Section §4.2 reviews prior work on evaluation methodologies for speech and audio models. Section §4.3 introduces SHALLOW, our framework for detecting and categorizing ASR hallucinations. Section §4.4 presents UnSLU-BENCH, our benchmark for evaluating machine unlearning in spoken language understanding. Section §4.5 describes ARCH, our unified benchmark for assessing audio representation learning. Finally, Section §4.6 concludes the chapter and outlines future research directions. A graphical taxonomy of the contributions presented in this chapter is given in Figure 4.1.

4.2 Related Work

The development of comprehensive evaluation frameworks is contingent on a clear understanding of existing methodologies and their limitations. This section reviews prior work in three key areas that correspond to the benchmarks introduced in this chapter: hallucination detection in speech (§4.2.1), machine unlearning for privacy (§4.2.2), and the benchmarking of general-purpose audio representations (§4.2.3).

4.2.1 Evaluating Hallucinations in Speech Models

The phenomenon of "hallucination," or the generation of content not grounded in a source input, is most widely studied in Natural Language Generation (NLG) and Large Language Models (LLMs), where it typically refers to the production of factually incorrect or fabricated information [162, 163]. Early metrics like BVSS were introduced to identify fluent yet nonsensical text using cosine similarity [164]. Much of the assessment in NLP relies on metrics like ROUGE, which is standard for summarization but requires a parallel corpus [165], or alternatives like GLEU that evaluate sentence-level fluency without this constraint [166]. In the visual domain, object hallucination in Large Vision-Language Models (LVLMs) describes the generation of incorrect image descriptions, such as mentioning objects that are not present [167].

In automatic speech recognition, however, hallucinations present a unique challenge. They manifest as fluent transcriptions that are completely unrelated to the acoustic input signal, often triggered by noise or ambiguity [168]. This problem arises from the need of the model to balance acoustic fidelity with linguistic coherence. The central issue is not factual inconsistency with world knowledge, but a lack of faithfulness to the spoken audio.

The standard metric for ASR, WER, calculates the edit distance between the hypothesis and a reference transcript [169]. While simple to compute, WER is a coarse measure that treats all errors equally, failing to distinguish between minor phonetic misrecognitions and critical semantic changes [170]. This limitation was recognized long before modern neural models, leading to the development of confidence measures designed to flag potentially incorrect words in a transcript [170–172]. However, these measures do not resolve WER's fundamental inability to capture semantic integrity or differentiate error severity [75, 173]. Other metrics have been proposed to address these shortcomings, such as Word Information Preserved (WIP) [174] for information transfer and embedding-based metrics for semantic similarity [75]. Yet, none of these are specifically designed to isolate hallucinations, which, as noted by [175], can be obscured by both high and low WER scores.

The urgency of this problem is amplified by the deployment of ASR in high-stakes fields like medicine and law. For instance, a study on the Whisper model found that while overall accuracy was high, about 1% of transcriptions contained fully

hallucinated phrases, with 38% of these fabrications promoting harmful or false content [11]. While detailed evaluation taxonomies exist for text generation (covering factual inconsistencies [176], knowledge conflicts [177], and attribution errors [178]) these are not directly applicable to ASR, which requires fidelity to an audio signal, not a textual source.

Recent work has begun to directly analyze ASR hallucinations. [179] treated them as general generative errors, while the authors of [7] used a perturbation-based method to measure the susceptibility of a model to hallucinate. Barański et al. [168] linked hallucinations to training data bias, and [175] proposed using an LLM to categorize different error types. However, these prior works remain limited in scope or rely on methodologies, like using an LLM, that are themselves prone to hallucination.

Our SHALLOW framework addresses this gap by providing the first systematic benchmarking framework to measure ASR hallucinations across distinct lexical, phonetic, morphological, and semantic dimensions.

4.2.2 Evaluating Machine Unlearning for Data Privacy

Beyond ensuring the fidelity of model outputs, a second equally important dimension of responsibility lies in safeguarding data privacy. Machine unlearning (MU) addresses this by providing a mechanism to remove the influence of specific training samples from a learned model without requiring a full and costly retraining process from scratch. This capability is not just a technical convenience; it is essential for complying with modern privacy regulations such as GDPR [80], which promotes the "*right to be forgotten.*" For speech data, which can contain sensitive biometric and personal information, MU is particularly crucial for building trustworthy AI systems.

The Machine Unlearning Problem

The task of machine unlearning can be formally defined as follows (refer to Figure 4.2 for a visual overview). Consider a model with parameters θ that has been trained on a dataset D. Each data point in this dataset is a triplet (x, y, s), where x represents the input data (such as an utterance), y is its corresponding label (such as an intent), and s identifies the data source (such as a specific speaker).

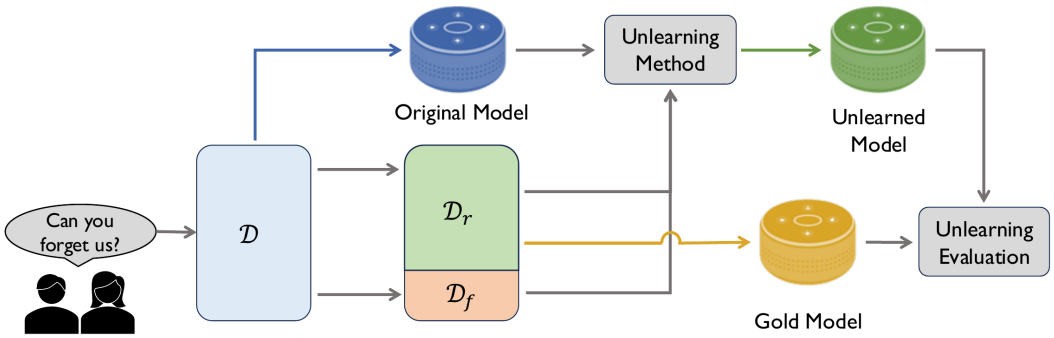


Fig. 4.2 **Machine Unlearning Pipeline in SLU setting.** Given a dataset $\mathscr{D}$ containing utterances annotated with intent and speaker identity, a subset of users requests data removal. This leads to a division into the retain set $\mathscr{D}_r$ and the forget set $\mathscr{D}_f$. The *Original* model is first trained on the complete dataset $\mathscr{D}$. An unlearning algorithm is then applied to produce an *Unlearned* model, aiming to erase the influence of $\mathscr{D}_f$ while preserving performance on $\mathscr{D}_r$. For reference, a `Gold` model is independently trained from scratch using only $\mathscr{D}_r$. The unlearned model is finally compared to the `Gold` model to evaluate the effectiveness of the unlearning process.

Let S be the set of all identities in D. If a subset of speakers, $\mathscr{S}_f \subset S$, requests the removal of their data, we define a "*forget set*" $D_f = \{(x,y,s) \in D \mid s \in \mathscr{S}_f\}$. The remaining data constitutes the "*retain set*," $D_r = D \setminus D_f$.

An unlearning process, denoted by a function $\phi(\cdot)$, takes the original model and the data partitions to produce a new, unlearned model: $\hat{\theta} = \phi(\theta, D_r, D_f)$ in which the influence of D_f has been erased. The objective for this new model, $\hat{\theta}$, is to approximate the behavior of a "*gold model*," θ', which is a model trained from the very beginning using only the retain set D_r, i.e., $\theta' = A(D_r)$, where A indicates the original training procedure. Since retraining θ' for every removal request is computationally prohibitive, the field of MU focuses on developing efficient, approximate unlearning procedures [180].

Core Evaluation Dimensions

The literature emphasizes that any MU method must be assessed along three fundamental dimensions: utility, efficacy, and efficiency [181].

Utility measures the ability of the model to maintain its performance on its original task after the unlearning process. It is typically quantified using metrics like accuracy or F1 score on a held-out test set. High utility is critical to ensure that the process

of forgetting does not lead to a catastrophic degradation of the model's overall capabilities [182].

Efficacy quantifies how completely the unlearned model, $\hat{\theta}$, has erased the influence of the forget set, D_f. This is often measured using Membership Inference Attacks (MIA) [181], which attempt to determine if a given data point was part of the model's training set. A lower MIA score on the forget set indicates better efficacy, as it suggests the model can no longer distinguish forgotten samples from unseen data, thus better approximating the privacy profile of the gold model. Deviations from this ideal, such as under-unlearning (incomplete removal) or over-unlearning (where forgotten data becomes anomalously distinct), can increase privacy risks [183].

Efficiency measures the computational cost of the unlearning procedure relative to the baseline of full retraining. For an unlearning method to be practical, it must offer significant savings in terms of time, computational resources, or energy consumption [9].

Composite Metrics

Evaluating these three dimensions in isolation is insufficient, as it can hide critical trade-offs. For example, a method could achieve perfect efficacy and efficiency by simply re-initializing the model's weights, but this would result in zero utility. This has led to the proposal of composite metrics. NoMUS [184], for example, combines efficacy and utility but overlooks the crucial aspect of efficiency. To provide a more holistic picture, we introduce the GUM metric as part of UnSLU-BENCH [26], the first to integrate all three pillars into a single score.

Machine Unlearning for Speech Models

While several progress has been made in MU techniques applied to computer vision and NLP [180, 185, 186], its application to speech processing is still in its early stages, and initial studies have been limited.

The authors of [187] studied machine unlearning in paralinguistic speech tasks, such as emotion recognition and depression detection. They focus on exact unlearning methods based on the SISA framework [188], where training is split into multiple shards to enable efficient removal of individual samples. These methods are

promising for performance retention and memory efficiency but are limited by their design: they require full control over the training pipeline, making them unsuitable for pre-trained or deployed models, and they do not scale well to large or dynamic environments.
Choi et al. [189] investigate MU in generative speech and music models, aiming to remove the influence of specific training songs in diffusion-based generators. While important for copyright protection, their goals differ from classification-oriented MU in speech understanding.
The work presented in [190] evaluates unlearning on keyword spotting and speaker identification with transformer models. However, the proposed benchmark is limited in linguistic and architectural diversity, focuses only on English, does not address SLU complexity, and compares only five unlearning techniques, providing a narrow coverage of existing methods.

To address these limitations, we created UnSLU-BENCH, the first systematic framework for evaluating MU specifically for spoken language understanding tasks. It establishes a foundation for assessing speaker-level unlearning across multiple languages and models, providing the standardized evaluation needed to advance responsible AI in speech.

4.2.3 Evaluating Audio Representation Learning

While the previous frameworks focus on specific behaviors of trained models, the ultimate quality of a foundation model lies in the generalizability of its learned representations. This requires a third type of evaluation framework dedicated to Audio Representation Learning (ARL). The central goal of ARL is to develop methods that encode raw audio signals into high-level, meaningful feature representations. These representations should be versatile enough to serve a wide array of downstream applications, including ASR [1], Music Information Retrieval (MIR) [191], and Acoustic Event Detection (AED) [192]. The rise of self-supervised learning approaches, exemplified by models like wav2Vec 2.0 [1] and HuBERT [2], has greatly advanced this area, particularly for speech-related tasks. These models learn powerful representations by pre-training on vast quantities of unlabeled audio data. However, their strong performance on speech has raised a critical question: *"How well do these speech-centric representations generalize to other audio domains, such as music or environmental sounds?"* Progress in ARL has been limited by

the scarcity of open-source models pre-trained on diverse, non-speech audio, which makes systematic evaluation challenging.

The growing demand for effective, general-purpose audio models has led to the creation of various evaluation frameworks [193–195]. However, a review of existing benchmarks reveals significant limitations in their scope, accessibility, or design philosophy. For instance, SUPERB [148] and LeBenchmark [196], while foundational for the speech community, are by design speech-specific and thus cannot be used to assess general audio understanding. HARES [197] broadens the scope but primarily evaluates representations extracted from spectrograms and is not fully open-source, limiting reproducibility. Other frameworks like LAPE [198] offer valuable insights but are tailored to specific research questions, such as performance in low-resource settings. The benchmark that most closely aligned with our goals is HEAR [199], which was designed to evaluate general-purpose ARL models across multiple domains. However, HEAR was created as a fixed, competition-style benchmark, which means it is not easily adaptable to new datasets or models as the field evolves. Furthermore, it shares very few datasets with other emerging frameworks, making cross-benchmark comparisons difficult. These limitations highlight the need for a new benchmark that is not only comprehensive in its domain coverage but also flexible and extensible by design.

We developed ARCH to address this gap. It provides a unified and modular platform for the standardized evaluation of ARL models across a diverse set of classification tasks, explicitly covering acoustic events, music, and speech. Unlike static benchmarks, ARCH is designed to be a community-driven resource that can readily incorporate new datasets and models, ensuring its long-term relevance. By offering a complementary evaluation perspective to existing work and emphasizing extensibility, ARCH facilitates a more complete and continuous assessment of progress in the quest for truly universal audio representation models.

4.3 SHALLOW: A Framework for Hallucination Detection

In this section, we introduce SHALLOW (**S**peech **HALL**ucination **O**vervie**W**), a novel evaluation framework specifically designed to detect, categorize, and measure

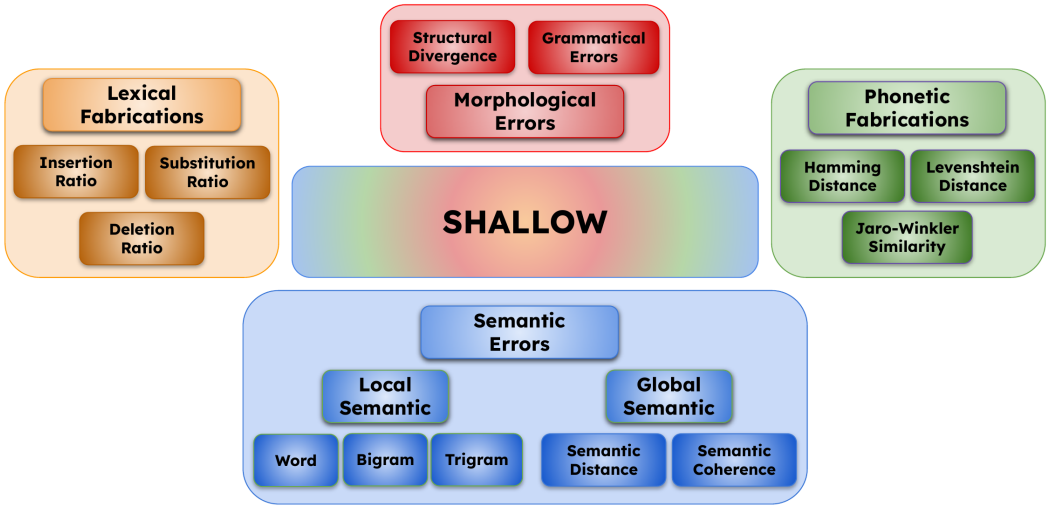


Fig. 4.3 **SHALLOW benchmark.** Overview of the proposed framework with its core dimensions: lexical fabrications, phonetic fabrications, morphological errors, and semantic errors.

hallucinations in ASR systems [25]. While WER treats all errors equally, SHALLOW is designed to recognize that different types of errors may have to significantly different impact on downstream applications and user experience. Consider two errors: changing "*the dog bit the man*" to "*the man bit the dog*", versus changing "*the dog is running*" to "*the dog runs*". While both might affect WER similarly, the first fundamentally alters meaning while the second preserves it. SHALLOW therefore provides a more nuanced assessment of ASR outputs, highlighting errors that are likely to have the greatest real-world impact.

4.3.1 Methodology

To capture these crucial differences, we evaluate ASR outputs through four complementary lenses: lexical accuracy (measuring word-level errors), phonetic similarity (assessing pronunciation-based mistakes), morphological correctness (analyzing grammatical structure), and semantic preservation (evaluating meaning consistency). Figure 4.3 shows an overview of our framework. Each hallucination axis is scored based on a mix of factors from linguistic and computational principles. We weigh these factors based on how much they matter in real-world deployments. This gives us a clearer and more detailed view of the reliability of a model than traditional accuracy scores can provide.

Lexical Fabrications

At its most basic level, ASR hallucinations manifest as incorrect words in the transcript. We track three distinct ways this can happen: the model can add words that were never spoken (insertions), replace actual words with incorrect ones (substitutions), or skip over spoken words entirely (deletions). Our scoring system pays special attention to insertions, as these represent pure fabrication rather than simple misrecognition:

$$LF = \begin{cases} 1 & \text{if } r_i = 1 \text{ AND } w_i \neq \text{fillers (e.g., 'uhm')} \\ 0.5 \cdot r_i + 0.3 \cdot r_s + 0.2 \cdot r_d & \text{otherwise} \end{cases} \tag{4.1}$$

Here, we calculate ratios for each error type: r_i for insertions, r_s for substitutions, and r_d for deletions, each divided by total word count. The weighting scheme (0.5, 0.3, 0.2) emerges from our analysis of real ASR errors across multiple datasets. Insertions receive the highest weight because they introduce content with no acoustic evidence. Substitutions get medium weight as they at least maintain some relationship to the audio signal. Deletions receive the lowest weight since omitting information, while problematic, typically causes less harm than fabricating new content. We recognize that different applications may place varying importance on each type of ASR error, so our framework allows these weights to be adjusted to match specific needs, ensuring that the evaluation reflects the practical impact of hallucinations in the target context.

Phonetic Fabrications

Sometimes ASR systems produce words that sound similar to what was actually said but are lexically incorrect. We capture these phonetic near-misses using three different distance measures. Hamming distance (H) counts mismatched characters, Levenshtein distance (L) counts minimum edits needed, and Jaro-Winkler similarity (JW) handles character transpositions and common prefixes. Each is normalized between 0 and 1, with higher values indicating greater divergence. We combine these into a single phonetic fabrication score:

$$PF = \frac{H_N + L_N + (1 - JW)}{3} \tag{4.2}$$

By using metaphone transformations [200] before computing these distances, we account for valid pronunciation variations. For example, "*recognize*" and "*recognise*" would score as identical, while "*recognize*" and "*wreck a nice*" would show high divergence despite similar phonetics.

Morphological Errors

Beyond simple word-level mistakes, ASR systems often make errors in grammar and sentence structure. These morphological errors might preserve the basic meaning but violate language rules. This becomes especially critical in formal settings like legal transcription, or when working with languages where small grammatical changes can significantly alter meaning. We break down morphological analysis into two key components: how sentences are structured and how grammar rules are followed.

Structural Divergence. To measure how sentence structure changes, we first convert both the reference and hypothesis into dependency graphs. These graphs show how words relate to each other, acting as maps of grammatical relationships. We then compute how different these graphs are using an inverse Jaccard similarity score. Mathematically, we have:

$$SD = 1 - \frac{|R \cap H|}{|R \cup H|} \tag{4.3}$$

Here, R and H denote the dependency relation sets extracted from the reference and hypothesis sentences, respectively.

Grammatical Errors. Not all grammar mistakes are equally serious. We use this weighted formula to reflect their different levels of importance:

$$GE = \frac{0.4 \cdot E_{Gr} + 0.3 \cdot E_{Sp} + 0.3 \cdot E_{Pu}}{N_{words}} \tag{4.4}$$

Grammar errors (E_{Gr}) get the highest weight (0.4) because they can completely change meaning: think, for example, of "*he is*" versus "*he are*". Spelling (E_{Sp}) and punctuation (E_{Pu}) errors share lower weights (0.3 each) since they typically cause less confusion.[1]

[1] While not all ASR systems output punctuation, we include it because punctuation errors often signal deeper problems with how the system understands sentence boundaries and structure. Full error definitions are at languagetool.org/rules

Overall Morphological Score. We combine structural and grammatical scores with carefully chosen weights:

$$ME = 0.4 \cdot SD + 0.6 \cdot GE \tag{4.5}$$

Grammar gets higher weight (0.6) because it more directly affects understanding. Structure (0.4) still matters but tends to cause fewer critical misunderstandings. Our experiments (detailed in [25]) validate these weightings across diverse speech samples.

Semantic Errors

The most difficult hallucinations to catch are those where the output is perfectly grammatical but means something different from what was said. Traditional metrics often miss these meaning-changing errors entirely. This becomes dangerous in critical applications. Imagine for example a medical transcription that changes "*increase the dose*" to "*decrease the dose*" while maintaining perfect grammar.

We look at semantic errors at two levels: local (within short phrases) and global (across entire utterances).

Local Semantic Errors. We examine meaning preservation using a sliding window approach at three scales. For each window size $w \in \{1,2,3\}$, we first convert text segments into semantic vector embeddings using a lightweight BERT [58]. We then compare each hypothesis window to all reference windows of the same size, and keep the best match score for each comparison. The local semantic error score combines these comparisons with decreasing weights for larger windows:

$$LS = 0.5 \cdot (1 - C_1) + 0.3 \cdot (1 - C_2) + 0.2 \cdot (1 - C_3) \tag{4.6}$$

Single words (C_1) receive the highest weight (0.5) since they form the core building blocks of meaning, directly reflecting word-level semantic changes. Bi-grams (C_2) and tri-grams (C_3) get lower weights (0.3 and 0.2) but help catch meaning shifts that only become apparent in longer sequences. This formulation prioritizes token-level distortions while remaining sensitive to broader semantic inconsistencies, allowing us to detect cases where individual words appear plausible on their own but produce contextual contradictions when combined.

Global Semantic Errors. Looking at small chunks of text is not enough, we also need to check if the overall meaning stays intact. We measure this through two complementary approaches: how far the meaning has drifted, and how well the logic holds together.

Semantic distance. First, we turn both the reference and hypothesis into dense vectors using a sentence embedding model [201]. These vectors capture the overall meaning of each sentence. We then measure how far apart these meanings are using cosine similarity and flip the score (1 minus similarity) to get a distance. A larger distance means the ASR output has drifted further from the intended meaning.

Semantic coherence. We combine two complementary tools to assess logical consistency. First, we use BERTScore [202] to measure semantic similarity between corresponding words. Next, we apply a natural language inference (NLI) model [203] to detect logical relationships between the reference and hypothesis. The NLI assigns different weights based on its prediction: (i) full coherence if the hypothesis logically follows from the reference (weight 1.0); (ii) partial coherence if it is neutral (0.5); and (iii) no coherence if it contradicts the reference (0.0). The final semantic coherence score is obtained by multiplying these two components, so high values occur only when both semantic similarity and logical consistency are preserved.

Overall Global Semantic Score. We weight distance and coherence metrics equally:

$$GS = \frac{(1 - SDist) + (1 - SC)}{2} \tag{4.7}$$

Our preliminary tests (detailed in [25]) show this balanced approach matches human judgments better than relying on either metric alone.

Aggregated Semantic Error Score. To create a final semantic score, we combine local and global measures with a careful weighting:

$$SE = \frac{1}{4} \cdot LS + \frac{3}{4} \cdot GS \tag{4.8}$$

The heavy emphasis on global scores (0.75 weight) reflects that overall meaning preservation usually matters more than local details. However, we maintain some weight (0.25) for local scores to catch subtle meaning shifts that might get lost in the bigger picture.

SHALLOW Evaluation Framework

Rather than aggregate everything into a single number, SHALLOW keeps track of four separate scores:

$$\text{SHALLOW} = \{LF, PF, ME, SE\}$$

Where *LF* tracks made-up words, *PF* measures sound-alike errors, *ME* catches grammar mistakes, and *SE* monitors meaning changes. This multi-score approach offers fine-grained insight by revealing specific model weaknesses that aggregate metrics often mask. By distinguishing between error types, researchers can focus on the most problematic aspects of performance for targeted optimization. The framework also aids practical decision-making, such as prioritizing semantic accuracy in safety-critical medical settings. Finally, this approach enables meaningful comparisons by highlighting qualitative differences overlooked by traditional metrics. While these automated metrics provide a robust evaluation, future work is still required to correlate these specific scores with human perception of error severity.

4.3.2 Experimental Setup

We designed our evaluation to test ASR hallucinations across diverse speaking conditions and model architectures.[2] Our selection of datasets covers common challenges in real-world ASR deployment, while our model choices represent major architectural approaches in modern speech recognition.

Dataset selection. As shown in Table 4.1, to capture different hallucination triggers, we organize our evaluation datasets into four categories.

Clean speech benchmarks. We start with widely-used datasets representing controlled recording conditions: LibriSpeech-Other [59] provides audiobook recordings with clean speech. TEDLIUM [204] offers prepared presentations with professional speakers. GIGASPEECH [205] adds variety through its multi-domain content.

Noisy environment data. CHiME-6 [206] presents a real-world challenge through dinner party recordings. These conversations include overlapping speech, back-

[2] github.com/SALT-Research/SHALLOW

Table 4.1 **SHALLOW evaluation datasets.** Statistics of the datasets used in the SHALLOW benchmark, divided by categories.

Dataset	# Test Utts	Domain	Characteristics
Standard Speech Conditions			
LibriSpeech (other) [59]	2,939	Read audiobooks	Standard "other" split with more challenging samples
TEDLIUM [204]	1,469	TED talks	Clear, prepared speech by professional speakers
GIGASPEECH [205]	25,619	Diverse sources	Audiobooks, podcasts, YouTube; diverse topics
Challenging Acoustic Environments			
CHiME-6 [206]	11,027	Dinner parties	Conversational speech with natural domestic noise
Heavily-Accented Domains			
CORAAL [207]	5,000	Interview speech	Regional varieties of African American Language
CV16-Accented [60]	2,197	Crowd-sourced	English utterances with accent variation
GLOBE-v2 [208]	5,046	Global accents	164 accents from worldwide speakers
SpeechOcean [209]	2,500	L2 English	Non-native speakers (L1: Mandarin); children and adults
Specialized Domains and Voices			
MyST Child [210]	13,180	Educational	Children (grades 3-5) with virtual science tutor
VoxPopuli [63]	1,842	Political speeches	Formal speaking with domain-specific terminology

ground noise, and room acoustics. This helps us understand how environmental factors trigger different types of hallucinations.

Accent variation. We examine accent-related hallucinations through four datasets: CORAAL [207] captures African American Language variations; CV16-Accented [60] provides diverse English accent patterns; GLOBE-v2 [208] covers 164 distinct English accents worldwide; and SpeechOcean [209] focuses on Mandarin speakers using English as a second language.

Domain-specific speech. Different speaking contexts bring unique vocabulary and patterns: MyST Child [210] contains children's speech in educational settings. VoxPopuli [63] provides political speeches with formal language. These datasets help identify domain-specific hallucination patterns.

Model selection. We evaluate four distinct families of ASR architectures to understand how design choices affect hallucination behavior. Table 4.2 summarizes the main characteristics of each of them.

Self-supervised speech encoders. HuBERT (`HuB`) [2] uses masked prediction to learn speech features. `MMS` [211] adds multilingual capability through training on 1,406 languages.

Encoder-decoder transformers. This family includes Whisper models (`WLv2` and `WLv3`) [13], which use weak supervision at scale. `Canary` [213] adds controlled

Table 4.2 **SHALLOW evaluation models.** Overview of the ASR models used in the SHALLOW benchmark, divided by categories.

Model	Architecture Type	# Params	Key Characteristics
Self-Supervised Speech Encoders			
HuBERT [2]	Encoder-only	300M	Masked prediction objectives; fine-tuned on LibriSpeech
MMS [211]	Encoder-only	1B	Multilingual (1,406 languages); language-agnostic representations
Encoder-Decoder Transformers			
Whisper-Large-v2 [212]	Encoder-decoder	1.5B	680,000 hours of weakly supervised multilingual training
Whisper-Large-v3	Encoder-decoder	1.5B	5M+ hours training data; enhanced generalization capabilities
Canary [213]	Encoder-decoder	1B	FastConformer encoder (32 layers); token-driven decoding
Encoder-Transducer Models			
Parakeet [72]	Encoder-transducer	1.1B	FastConformer-based; optimized for English recognition
Multimodal SpeechLLMs			
SALMONN [100]	Decoder w/ encoders	7B	Integrates LLMs with speech/audio encoders; unified processing
Qwen2Audio [101]	Decoder w/ encoders	8.4B	Part of Qwen2 series; specialized audio encoders
Qwen2.5-Omni [107]	Decoder w/ encoders	10.7B	Enhanced Qwen2; broader multimodal capabilities
Granite-Speech [214]	Decoder w/ encoders	8.6B	Two-pass design for transcription and translation
Kimi-Audio [215]	Decoder w/ encoders	9.7B	Open audio model; unified framework for audio tasks
Phi4-MM-Instruct [102]	Decoder w/ encoders	5.6B	Open-weights foundation model; Multimodal by design.

text generation through token-driven decoding. These models balance acoustic and linguistic processing through separate encoder-decoder components.

Encoder-transducer models. `Parakeet` [72] represents this family, using FastConformer architecture [73] for tight audio-text alignment. This design creates direct connections between acoustic and linguistic features.

Multimodal SpeechLLMs. This newest category integrates speech processing into language models: SALMONN (`SALM.`) [100], Qwen2Audio (`Q2A`) [101], Qwen2.5Omni (`Q2.5O`) [107], Granite-Speech (`Granite`) [214], Kimi-Audio (`Kimi`) [215], and Phi4-Multimodal-Instruct (PHI4) [102].

Table 4.3 **SHALLOW synthetic data, examples.** Overview of synthetic data categories, with examples, description, WER and SHALLOW metrics.

Category	Description	Reference	Hypothesis	WER	LF	PF	ME	SE
Lexical	Adds unrelated or hallucinated words	They are playing chess outside	They are playing chess outside with magical stones	0.60	0.19	0.31	0.15	0.29
Phonetic	Substitutes with phonetically similar but incorrect words	I went to the retirement party	I bent to the retirement party	0.17	0.05	0.04	0.27	0.13
Morphological	Tense or agreement errors	They sing together every morning	They sings together every mornings	0.40	0.12	0.02	0.40	0.08
(Local) Semantic	Replaces a single word, changing the meaning	He painted the fence	He destroyed the fence	0.25	0.08	0.37	0.34	0.66
(Global) Semantic	Changes sentence meaning	They went to the beach for vacation	They stayed home for vacation	0.57	0.14	0.45	0.36	0.68
Mixed Errors	Combines lexical, morph. and semantic hallucinations	She fixed her broken glasses	She fix broken lens with dragon spark	1.20	0.38	0.51	0.40	0.39
WER only	High WER, same meaning	They joined us for dinner	They came over to eat with us	1.20	0.38	0.64	0.40	0.17

We evaluate all models using their original pre-trained weights without domain adaptation. This approach reveals their inherent tendencies toward hallucination.

4.3.3 Results and Discussion

Benchmark Validation

We first validate SHALLOW's ability to identify distinct types of hallucinations beyond what WER can measure. Using GPT-4o [216], we create a controlled synthetic dataset designed to test each error type independently.

The validation dataset contains 1,050 reference-hypothesis pairs spread across six categories of 150 samples each: lexical, phonetic, morphological, local semantic, global semantic errors, and mixed error types. Each sample maximizes one type of hallucination while minimizing others, allowing precise evaluation of each SHALLOW metric. Examples of this datasets are shown in Table 4.3.

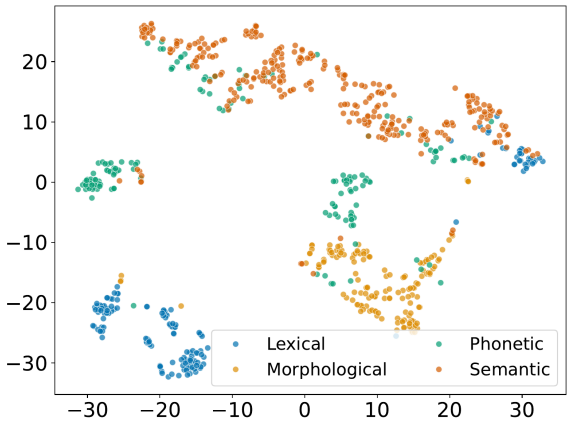


Fig. 4.4 **SHALLOW, t-SNE projection.** SHALLOW metrics on synthetic data.

Figure 4.4 visualizes these samples using t-SNE projection of their SHALLOW metric vectors. The clear clustering pattern confirms that our metrics successfully distinguish different types of hallucinations. Lexical and morphological errors form

Table 4.4 **SHALLOW and WER metrics.** Average scores across all datasets, per model. Best models are in **bold** (lower is better).

	HuB	MMS	WLv2	Canary	WLv3	Parakeet	SALM.	Q2A	Granite	Kimi	Q2.5O	Phi4
WER	40.94	27.45	19.12	14.26	14.20	12.54	99.92	21.99	15.21	13.53	12.76	**12.07**
Lexical	14.56	11.03	8.08	5.43	6.74	5.38	13.59	7.13	5.56	6.92	**5.17**	6.18
Phonetic	35.56	26.94	20.38	16.14	17.75	**15.33**	27.90	21.82	15.80	20.45	16.25	17.94
Morph.	27.55	23.54	13.15	11.05	11.13	10.59	16.54	13.77	**10.13**	12.30	10.56	11.22
Semantic	35.30	26.11	17.37	14.98	14.74	13.33	23.23	19.55	13.56	15.48	**12.71**	14.37

tight, separate clusters in the projection space. Phonetic errors show some overlap due to surface-level similarities between error types. Semantic errors appear more spread out, reflecting their contextual complexity. This clustering validates that SHALLOW metrics capture distinct aspects of hallucination behavior that WER cannot detect.

Analysis of Model Performance

Table 4.4 reveals important patterns that WER alone misses across all real-world datasets. While decoder-only models like Phi4 and Qwen2.5Omni achieve the lowest WER scores, SHALLOW exposes more nuanced behavior.

Parakeet excels at phonetic accuracy and ranks second in morphological correctness, matching its encoder-transducer architecture's focus on acoustic modeling. Qwen2.5Omni shows strength in lexical and semantic metrics, likely due to its robust language modeling. Models with similar WER, such as Whisper Large-v3 and Canary, show different hallucination patterns: Whisper produces more semantically coherent output, while Canary makes fewer lexical mistakes.

Notably, SALMONN demonstrates that high WER doesn't necessarily predict hallucination behavior. Despite poor overall accuracy, its lexical and semantic scores remain moderate. This finding confirms the value of SHALLOW in diagnosing specific failure modes even in lower-performing models.

Dataset-specific findings. Across different datasets (Table 4.5), we observe distinct patterns of hallucination behavior.

The CORAAL dataset shows increased hallucination metrics across all models, with SpeechLLMs particularly struggling compared to other architectures. This pattern

Table 4.5 **Evaluation of SHALLOW and WER metrics across all datasets.** Performance comparison of WER and SHALLOW metrics over all dataset categories. Dataset types are represented as: Standard Speech, Challenging Acoustic, Heavily-Accented, Specialized Domains, and AVG (overall average). The best results within each dataset are underlined, while the best overall averages are shown in **bold**.

Dataset	Metrics	Models											
		HuB	MMS	W-Lv2	Canary	W-Lv3	Parakeet	SALM.	Q2A	Granite	Kimi	Q2.5O	Phi4
CHiME-6	WER	59.41	57.30	32.43	34.16	30.25	29.23	136.93	30.93	41.08	33.59	29.92	29.42
	LF	24.20	24.46	15.16	13.20	14.76	13.80	18.53	11.27	13.84	17.90	13.56	15.3
	PF	53.39	55.66	33.20	32.76	30.89	33.49	38.85	33.40	33.41	42.84	32.92	37.36
	ME	37.32	40.27	18.34	19.00	17.28	20.01	22.10	18.46	18.44	23.30	19.04	21.29
	SE	48.02	51.30	26.88	29.45	25.17	27.78	32.31	27.79	27.15	32.83	25.26	30.43
CORAAL	WER	45.05	52.74	22.85	16.58	19.96	22.47	75.08	27.34	22.56	24.16	22.89	23.67
	LF	15.82	19.32	12.94	7.77	10.20	9.56	12.31	7.49	8.79	12.02	8.31	10.39
	PF	40.57	44.55	28.19	21.24	24.49	25.59	29.14	25.73	25.23	35.22	27.23	30.64
	ME	32.35	37.88	17.11	14.01	14.68	17.33	17.54	16.26	15.22	19.53	16.24	18.14
	SE	36.35	44.30	23.08	18.28	19.78	21.12	23.08	20.07	20.43	25.69	20.63	24.95
CV16-Accent	WER	96.02	18.43	20.56	8.08	11.37	5.71	46.26	90.30	6.28	6.87	6.30	6.52
	LF	29.85	5.70	4.23	2.50	3.11	1.71	10.72	26.3	1.93	2.23	2.02	2.09
	PF	69.06	11.75	10.49	6.63	8.12	4.43	43.42	59.95	5.45	6.10	5.61	5.66
	ME	51.16	18.67	9.97	8.39	8.80	6.13	22.24	38.17	6.04	6.60	6.07	6.73
	SE	79.05	16.39	11.55	8.80	9.52	5.58	39.71	68.30	6.56	6.56	6.40	6.36
GigaSpeech	WER	21.13	22.95	15.52	13.79	13.71	11.37	71.62	11.93	18.85	12.64	12.35	12.39
	LF	10.61	14.58	13.91	5.31	13.41	6.37	12.77	5.26	7.38	13.28	7.06	13.02
	PF	26.31	34.87	31.44	16.12	29.16	16.88	27.47	16.36	17.55	31.77	19.65	30.43
	ME	19.77	24.71	16.53	10.15	15.62	10.55	15.91	10.09	10.69	16.61	11.86	16.31
	SE	21.42	27.88	22.95	13.04	22.28	12.81	22.31	12.32	14.59	23.64	13.67	21.49
GLOBE-v2	WER	96.01	12.66	2.89	3.25	1.57	1.17	3.66	4.92	1.47	2.09	3.28	2.68
	LF	30.13	4.42	0.95	1.17	0.58	0.46	1.27	1.61	0.54	0.74	1.02	1.01
	PF	66.80	9.4	2.89	3.39	2.00	1.24	4.34	6.68	1.94	3.54	4.69	3.52
	ME	52.76	14.23	2.73	3.76	1.96	1.50	4.08	5.19	1.73	2.34	3.68	3.11
	SE	78.55	11.91	2.87	3.84	1.87	1.18	4.34	4.79	1.74	2.25	2.84	3.03
LibriSpeech	WER	3.51	7.95	6.15	3.88	3.98	2.62	4.94	3.98	2.98	2.75	3.46	3.83
	LF	1.29	2.88	2.45	1.48	1.48	1.00	1.89	1.42	1.13	1.16	1.35	1.56
	PF	4.48	8.7	7.87	5.31	5.24	3.54	7.46	5.38	4.46	4.35	5.22	5.47
	ME	5.68	11.44	7.58	5.92	5.55	4.35	7.09	5.71	4.47	4.42	5.33	5.80
	SE	3.51	8.81	7.19	5.02	4.63	2.95	6.60	4.58	3.61	3.40	3.96	4.88
MyST	WER	21.98	28.72	20.3	20.99	19.33	13.38	34.46	18.28	18.29	17.64	20.96	14.31
	LF	9.11	12.09	6.89	6.16	6.78	5.61	7.38	5.33	5.79	7.45	6.52	6.54
	PF	24.84	29.42	20.38	22.72	20.19	17.33	20.28	18.76	17.86	22.95	21.63	18.73
	ME	20.45	25.33	12.37	13.34	12.34	11.65	13.18	12.36	11.54	15.00	13.80	12.30
	SE	19.35	26.6	13.97	19.20	13.84	12.50	15.14	13.58	12.63	14.83	14.28	13.37
SpeechOcean	WER	37.98	47.04	25.37	25.35	21.16	23.90	25.98	15.66	24.70	19.92	13.48	12.88
	LF	12.98	15.45	8.19	8.99	7.46	8.14	7.04	5.02	8.41	6.15	4.43	4.27
	PF	21.37	27.75	16.77	17.24	16.17	16.76	15.29	13.64	16.18	15.21	9.83	9.32
	ME	25.30	32.89	15.28	17.15	14.69	16.67	14.88	12.20	15.3	14.59	10.95	10.92
	SE	31.04	41.11	22.43	24.92	21.31	23.69	20.81	16.86	22.34	18.37	14.26	13.76
TEDLIUM	WER	14.26	17.91	18.22	10.29	10.06	10.17	591.01	9.41	10.24	8.12	9.18	9.13
	LF	7.04	8.59	6.81	5.66	6.28	5.37	61.22	5.40	5.93	5.71	5.61	5.75
	PF	27.34	31.62	24.24	22.96	23.91	22.42	75.70	23.97	23.90	25.99	23.30	25.87
	ME	16.88	20.11	12.00	12.25	11.45	11.87	40.24	12.17	11.85	13.02	12.63	11.50
	SE	25.00	26.23	20.73	22.42	20.75	21.62	61.35	21.95	21.99	21.45	21.32	20.99
VoxPopuli	WER	14.05	8.75	26.92	6.22	10.57	5.42	9.21	7.17	5.65	7.53	5.77	5.86
	LF	4.59	2.77	9.28	2.02	3.30	1.75	2.79	2.15	1.86	2.54	1.83	1.90
	PF	21.48	15.66	28.37	12.99	17.29	11.59	17.04	14.30	12.05	16.52	12.45	12.37
	ME	13.88	9.88	19.54	6.55	8.97	5.80	8.14	7.06	6.02	7.57	5.98	6.09
	SE	10.69	6.54	22.09	4.88	8.22	4.11	6.64	5.25	4.60	5.77	4.47	4.41
AVG	WER	40.94	27.45	19.12	14.26	14.20	12.54	99.92	21.99	15.21	13.53	12.76	**12.07**
	LF	14.56	11.02	8.08	5.43	6.74	5.38	13.59	7.13	5.56	6.92	**5.17**	6.18
	PF	35.56	26.94	20.38	16.14	17.75	**15.33**	27.90	21.82	15.80	20.45	16.25	17.94
	ME	27.56	23.54	13.15	11.05	11.13	10.59	16.54	13.77	**10.13**	12.29	10.56	11.22
	SE	35.29	26.11	17.37	14.99	14.74	13.33	23.23	19.55	13.56	15.48	**12.71**	14.37

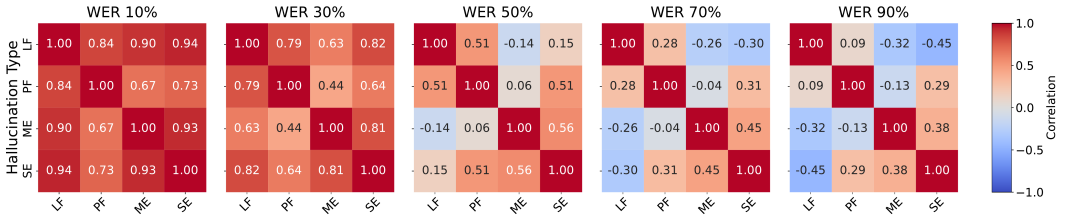


Fig. 4.5 **Spearman ρ.** Correlation between SHALLOW metrics and WER across varying error levels.

reveals the challenges of linguistic mismatch and emphasizes the need for more inclusive acoustic-linguistic modeling approaches.

In CHiME-6, phonetic hallucination scores remain consistently high regardless of model architecture. This uniform elevation in phonetic errors suggests that conversational overlap and acoustic degradation create fundamental challenges for phoneme-level decoding, persisting even when overall WER appears acceptable. SHALLOW metrics make this specific failure mode visible even when aggregate performance metrics might suggest adequate transcription quality.

In contrast, simpler datasets like LibriSpeech show consistently low hallucination scores across all dimensions. This alignment between SHALLOW metrics and expected dataset difficulty provides additional validation of our metrics' interpretive value.

Metric correlations. Figure 4.5 presents the Spearman correlation between the four hallucination metrics in SHALLOW across different WER ranges. At low WER (10-30%), the metrics show strong monotonic relationships ($\rho_s > 0.80$), suggesting that when errors are rare, they tend to manifest similarly across all categories.

However, these correlations weaken substantially as WER increases. By 50% WER, several metric pairs show weak or negative associations, such as the Lexical-Morphological correlation dropping to $\rho_s = -0.14$. At 90% WER, correlations deteriorate further, with Lexical-Semantic falling below -0.45.

This decoupling of hallucination types under degraded conditions reveals an important pattern: models may generate syntactically correct but semantically nonsensical outputs, or preserve meaning while distorting surface forms. These findings validate SHALLOW's multidimensional design for ASR evaluation. Where WER obscures error patterns, especially in challenging conditions, SHALLOW metrics remain discriminative and interpretable, revealing distinct error behaviors that emerge as

Table 4.6 **Medical ASR case study.** Evaluation of the Phi4 model on clinical speech data using SHALLOW metrics for detailed error analysis.

Reference	Hypothesis	WER	LF	PF	ME	SE
Medical-ASR Dataset						
i can not rotate my neck	i can rotate my neck	0.16	3.33	29.36	6.67	60.79
i feel like the room is spinning	i feel like the room is empty	0.14	4.29	18.14	29.09	56.75
is my cut infected or just healing	is my cat infected or just healing	0.14	4.29	0.00	17.78	56.27
i have a problem in my back i cannot extend it	i have a problem in my bag i cannot stand it	0.18	5.45	9.63	29.47	60.46
it is hard to see things	it is hard to say things	0.17	5.00	0.00	26.67	60.11
i feel pain in my knee	i feel pain in my neck	0.17	5.00	9.33	26.67	51.55
i feel lightheaded	i feel light headed	0.67	22.50	26.80	27.00	8.51
i cant breathe	i can not breathe	0.67	22.50	29.22	26.67	11.16
red flushes accompanied with itchy	red flush is accompanied with itching	0.60	20.33	33.64	36.00	13.27
AfriSpeech Dataset (Clinical Domain)						
the ulna remains relatively stationary	the owner remains relatively stationary	0.20	6.00	12.48	22.86	37.63
took 62 and 35 cc well with yellow nipple	took 62 and 335 cc well with yellow nipple	0.11	3.33	0.00	8.00	42.44
reason bilat pe eval for dvt	reason bilateral p e evaluation for dvt	0.67	22.00	32.02	42.67	19.38

recognition quality deteriorates. This capability proves particularly valuable in low-resource, noisy, or out-of-domain settings, where model hallucination profiles can diverge dramatically despite similar overall error rates.

Case study: Medical ASR

To demonstrate the practical value of SHALLOW in critical applications, we conduct a zero-shot analysis of Phi4 ASR performance in medical settings using the Medical-ASR[3] and AfriSpeech [217] clinical datasets. Table 4.6 presents cases where WER fails to capture potentially dangerous transcription errors.

Consider the transcription of "*I can not rotate my neck*" as "*I can rotate my neck*" (first row). While this error produces a falsely low WER of 0.16, SHALLOW reveals its severity through a high semantic error score ($SE = 60.79$), correctly flagging the critical inversion of the patient's reported symptom.

Similarly, when "*I feel like the room is spinning*" becomes "*I feel like the room is empty*," a crucial indicator of vertigo disappears. Despite the low WER (0.14), SHALLOW's high semantic error score (56.75) appropriately flags the loss of vital diagnostic information.

Even single-word substitutions can have serious implications. Changing "*cut*" to "*cat*" in an infection-related query yields a low WER (0.14) but fundamentally alters

[3] https://huggingface.co/datasets/jarvisx17/Medical-ASR-EN

the medical context. SHALLOW captures this through elevated semantic error scores (56.27) despite low phonetic fabrication values.

In the AfriSpeech dataset, the transformation of "*the ulna remains relatively stationary*" to "*the owner remains relatively stationary*" demonstrates how phonetically plausible errors ($PF = 12.48$) can produce medically nonsensical statements. SHALLOW's semantic error score (37.63) correctly identifies this problem despite the low WER (0.20).

These examples highlight the ability of SHALLOW to identify potentially harmful transcription errors that traditional metrics might overlook, making it particularly valuable for evaluating ASR systems in healthcare applications.

4.3.4 Summary and Practical Implications

SHALLOW establishes a principled foundation for evaluating hallucinations in automatic speech recognition, moving beyond aggregate word-level accuracy toward multidimensional behavioral assessment. By decomposing hallucinations into lexical, phonetic, morphological, and semantic components, the framework provides interpretable, application-aware insight into how ASR systems fail. Our experiments reveal that hallucination profiles diverge sharply across architectures and acoustic conditions, even when WER remains similar, underscoring the inadequacy of traditional single-score metrics for diagnosing reliability in complex or high-stakes domains.

From a practical perspective, SHALLOW offers several concrete advantages for ASR development and deployment. First, it exposes the underlying structure of model errors, enabling targeted improvement instead of broad retraining. Second, it provides a standardized diagnostic layer to monitor hallucination behavior across updates, datasets, or fine-tuning stages. Third, it introduces a transparent and interpretable evaluation signal that can inform model selection and regulatory assessment, especially in domains such as healthcare, education, or accessibility technology.

The medical ASR case study highlights SHALLOW's operational value: despite low WER, models frequently produced meaning-altering transcriptions that could change diagnostic interpretation. By capturing such errors through elevated semantic scores, SHALLOW directly addresses the gap between statistical correctness and communicative reliability. This makes it not just a benchmarking tool but a risk-

awareness framework, capable of identifying when ASR outputs cross from benign inaccuracies into potentially harmful hallucinations.

Ultimately, SHALLOW reframes ASR evaluation from "*how accurate is the output?*" to "*how trustworthy is the model under real-world variability?*". This shift enables safer, more accountable deployment of speech recognition systems and encourages model designs that prioritize robustness, interpretability, and human-centered reliability.

4.4 UnSLU-BENCH: Machine Unlearning for SLU

Modern privacy regulations increasingly require the ability to remove specific user data from trained models. This "*right to be forgotten*" presents unique challenges for speech models, which may encode sensitive speaker information across multiple layers. We present UnSLU-BENCH [26], the first comprehensive **BENCH**mark for evaluating machine **Un**learning in **S**poken **L**anguage **U**nderstanding models. Our framework assesses both the effectiveness of forgetting and its impact on model performance.

4.4.1 Methodology

Unlearning Methods

We benchmark eight distinct approaches to machine unlearning, each offering different trade-offs between forgetting effectiveness and model utility.

Fine-Tuning (`FT`) serves as our baseline approach. It continues training the model for one epoch using only the retained data D_r. The intuition is that the forgotten data D_f becomes less influential through this additional training pass.

Negative Gradients (`NG`) [180] takes a more direct approach to forgetting. It fine-tunes the model exclusively on D_f but reverses the gradient direction during backpropagation. This actively pushes the model away from knowledge gained from the forgotten data.

NegGrad+ (`NG+`) [184, 218] extends the negative gradient approach. It combines reversed gradients on D_f with standard fine-tuning on D_r. This balanced approach

helps prevent catastrophic forgetting, where removing specific knowledge damages the model's overall capabilities.

Catastrophic Forgetting (`CF-k`) [219] takes a targeted approach to unlearning. It applies fine-tuning only to the final k layers of the model, where task-specific representations concentrate. This selective update increases efficiency while maintaining lower-level feature extractors.

UNSIR (`UNSIR`) [9] implements a two-phase strategy. The impair phase creates error-maximizing noise for each sample in D_f and trains the model with this noise. The repair phase then rebuilds model utility through standard fine-tuning.

Bad Teaching (`BT`) [220] employs a knowledge distillation framework with two teachers. A competent teacher (copy of original model) guides learning on D_r. An incompetent teacher (untrained model) guides behavior on D_f. We also evaluate a light variant (`BT-L`) that replaces the incompetent teacher with random predictions.

SCRUB (`SCRUB`) [218] uses a single-teacher distillation setup. It combines three objectives: maximizing teacher similarity on D_r, minimizing it on D_f, and maintaining task performance.

Evaluation Framework

Effective unlearning requires balancing three fundamental aspects: *efficacy* (successful forgetting), *efficiency* (computational cost), and *utility* (preserved performance). Ignoring any dimension leads to trivial solutions. Without measuring efficacy, keeping the original model unchanged appears optimal. Without considering efficiency, complete retraining becomes the default choice. Without tracking utility, random predictions would suffice.

We introduce the *Global Unlearning Metric (GUM)* to comprehensively evaluate unlearning success across all three dimensions. Each component is carefully designed to capture specific aspects of the unlearning process, with all scores normalized to the range $[0,1]$ for consistent comparison.

Utility score. The utility score measures how well the unlearned model maintains its original task performance compared to the gold standard:

$$U = 1 - |F1_T^{(g)} - F1_T^{(u)}| \tag{4.9}$$

where $F1_T^{(g)}$ is the macro F1 score of the gold model on the test set and $F1_T^{(u)}$ is the macro F1 score of the unlearned model. The absolute difference measures performance deviation, subtracting from 1 further converts deviation to similarity. A score of 1 indicates identical performance to the gold model, while 0 indicates complete performance degradation.

Efficacy score. The efficacy score quantifies how successfully the model has forgotten the target data through membership inference attack (MIA) resistance:

$$E = 1 - \left(\frac{\text{MIA}'^{(u)} - \text{MIA}'^{(g)}}{\text{MIA}^{(o)} - \text{MIA}'^{(g)}} \right)^2 \tag{4.10}$$

Here, $\text{MIA}^{(o)}$ is the MIA success rate on the original model, $\text{MIA}^{(g)}$ is the MIA success rate on the gold model (ideal target), $\text{MIA}^{(u)}$ is the MIA success rate on the unlearned model, $\text{MIA}'^{(u)} = \min\{\text{MIA}^{(u)}, \text{MIA}^{(o)}\}$ prevents scores exceeding original model, $\text{MIA}'^{(g)} = \min\{\text{MIA}^{(g)}, (\text{MIA}'^{(u)} + \text{MIA}^{(o)})/2\}$ handles edge cases. The squared term emphasizes small differences between unlearned and gold model MIA rates. Perfect efficacy ($E = 1$) means the unlearned model matches the MIA resistance of the gold model.

Efficiency score. The efficiency score evaluates computational cost relative to complete retraining:

$$T = 1 - \frac{log(T^{(u)} + 1)}{log(T^{(g)} + 1)} \tag{4.11}$$

where $T^{(u)}$ is the time required for unlearning and $T^{(g)}$ is the time required for complete retraining. The logarithmic scaling handles large time differences, adding 1 prevents undefined values for instant operations. A score of 1 indicates instant unlearning, while 0 indicates time equal to or exceeding complete retraining.

Global Unlearning Metric. The final GUM score combines all three components through a weighted harmonic mean:

$$GUM = \frac{(1 + \alpha + \beta)UET}{\alpha ET + \beta UT + UE} \tag{4.12}$$

Here, α weights the relative importance of efficacy, β weights the relative importance of efficiency. The harmonic mean ensures all components must be reasonable for a good score; poor performance in any dimension significantly impacts the final score.

In our benchmark evaluation, we set $\alpha = \beta = 1$ to weight all aspects equally. This balanced weighting reflects our view that successful unlearning requires satisfactory performance across all three dimensions. The harmonic mean formulation ensures that methods cannot achieve high GUM scores by excelling in one aspect while failing in others.

A GUM score of 1.0 represents the ideal case: perfect utility preservation, complete forgetting, and instant computation. Scores typically fall well below this ideal, reflecting the inherent trade-offs in machine unlearning. The metric provides a single, interpretable value for comparing different unlearning approaches while capturing the multi-faceted nature of the task.

4.4.2 Experimental Setup

Dataset selection. Our benchmark evaluates unlearning across four intent classification datasets representing different languages and complexity levels.[4] The FSC [146] dataset provides a controlled evaluation environment with 31 distinct intents in English. SLURP [147], ITALIC [30], and SpeechMASSIVE [221] offer more challenging scenarios with 60 intents each and greater linguistic complexity. ITALIC extends SLURP's task structure to Italian, while SpeechMASSIVE covers multiple languages, though we focus specifically on German and French.

The original SLURP dataset lacks speaker-independent splits, which are crucial for effective unlearning evaluation. We address this by creating new splits that ensure speaker separation between retain, forget, and test sets[5]. We refer to this modified version as SLURP*.

For all other datasets, we maintain their original speaker-independent splits. To create forget sets that realistically simulate user deletion requests, we randomly select speakers who have contributed at least 100 audio samples. This threshold ensures sufficient representation in the original training. The resulting forget sets comprise 2.5-5% of the total data, matching realistic scenarios where individual users request data removal.

Model selection. We evaluate two model architectures for each language setting. For English datasets, we use the base versions of wav2vec 2.0 [1] and HuBERT [2].

[4] github.com/koudounasalkis/UnSLU-BENCH

[5] These splits are publicly available in our project repository.

Table 4.7 **Unlearning results on FSC.** $F1_T$ represents the macro F1 score on the test set, and $F1_F$ the macro F1 on the forget set. Best results (i.e., values closest to the gold model for F1 and MIA, and highest for all other metrics) are shown in **bold**, with second-best results underlined. Original and gold model scores are highlighted for reference.

Method	FSC									
	wav2vec 2.0					HuBERT				
	$F1_T$	**$F1_F$**	**MIA**	**GUM**	**Speedup**	**$F1_T$**	**$F1_F$**	**MIA**	**GUM**	**Speedup**
Orig.	.994	1.00	.508	.000	1.00×	.993	1.00	.511	.000	1.00×
Gold	.993	.997	.503	.000	1.00×	.991	.996	.507	.000	1.00×
FT	**.993**	.999	**.504**	.517	7.960×	.979	.993	**.508**	.514	7.690×
NG	.987	.976	.501	**.816**	**206.9**×	**.992**	**.996**	.514	.000	**201.1**×
NG+	.994	.994	.493	.000	4.030×	.979	.929	.510	.336	3.900×
CF-k	.994	1.00	.501	.606	16.97×	.993	1.00	.505	**.642**	26.70×
UNSIR	.991	1.00	.506	.447	6.550×	.994	.998	**.508**	.484	6.380×
BT	**.993**	1.00	.508	.000	4.780×	.993	.999	.504	.363	4.650×
BT-L	.994	**.996**	.506	.431	5.870×	.993	.997	**.506**	.464	5.690×
SCRUB	.994	1.00	.506	.439	6.210×	.993	.998	**.508**	.479	6.220×

For multilingual datasets, we employ XLS-R 128 [14] and XLS-R 53 [222], the latter with language-specific ASR fine-tuning.

Unlearning configuration. We group unlearning methods into two categories based on their potential impact on model performance. For methods that risk significant disruption (NG, NG+, BT, BT-L, SCRUB), we use conservative learning rates: $5e-07$, $1e-06$, and $5e-06$. For more stable methods (FT, CF-k, UNSIR), we use larger learning rates: $1e-05$, $5e-05$, and $1e-04$. We select the best configuration for each method based on our GUM metric, which balances utility, efficacy, and efficiency. For UNSIR, which was originally designed to forget entire classes, we adopt the sample-level adaptation proposed by [184].

4.4.3 Results and Discussion

Our experiments analyze machine unlearning behavior across different SLU models and datasets, revealing several key patterns.

Comprehensive benchmark analysis. Tables 4.7, 4.8, and 4.9 demonstrate distinct performance patterns across unlearning methods over the considered datasets. We evaluate both F1 scores and MIA results against the gold model as our target baseline.

Table 4.8 **Unlearning results on SLURP* and ITALIC.** Best results are shown in **bold**, with second-best results underlined. Original and gold model scores are highlighted for reference.

Method	SLURP*										ITALIC									
	wav2vec 2.0					HuBERT					XLS-R 128					XLS-R 53-IT				
	$F1_T$	$F1_F$	MIA	GUM	Speedup	$F1_T$	$F1_F$	MIA	GUM	Speedup	$F1_T$	$F1_F$	MIA	GUM	Speedup	$F1_T$	$F1_F$	MIA	GUM	Speedup
Orig.	.689	1.000	.628	.000	1.000×	.712	1.000	.613	.000	1.000×	.688	.894	.632	.000	1.000×	.778	1.000	.615	.000	1.000×
Gold	.707	.711	.506	.000	1.000×	.704	.715	.492	.000	1.000×	.643	.568	.532	.000	1.000×	.784	.736	.478	.000	1.000×
FT	.638	**.970**	.648	.000	83.78×	.734	1.000	.611	.088	79.00×	**.638**	.671	.555	.590	30.80×	.711	.850	.550	.551	31.10×
NG	.695	.986	.604	**.563**	**1748**×	.718	.959	.587	**.587**	**1654**×	.679	.868	.603	**.646**	**613.4**×	.590	.621	**.525**	**.766**	**623.0**×
NG+	.701	.995	**.603**	.446	41.63×	.630	**.852**	**.453**	.578	39.30×	.658	.001	.932	.000	15.14×	.743	.936	.582	.418	15.37×
CF-k	**.709**	1.000	.626	.089	291.9×	.715	1.000	.608	.196	274.2×	.677	.871	.626	.253	98.59×	**.781**	1.000	.609	.201	98.99×
UNSIR	.673	1.000	.637	.000	64.07×	.722	1.000	.613	.000	60.44×	.636	.830	.621	.328	22.01×	.775	1.000	.612	.109	22.26×
BT	.710	.999	.619	.275	50.35×	.711	1.000	.613	.000	47.42×	.683	**.639**	.481	.504	17.90×	.731	**.848**	.557	.491	17.94×
BT-L	.680	.995	.637	.000	61.74×	.685	.907	.558	.578	58.11×	.686	.651	.518	.558	22.02×	.729	.876	.564	.499	22.21×
SCRUB	.697	.999	.608	.429	64.82×	**.704**	1.000	.600	.350	65.40×	.442	.357	**.533**	.536	23.25×	.770	.990	.610	.164	22.66×

Table 4.9 **Unlearning results on SpeechMASSIVE de-DE and fr-FR.** Best results are shown in **bold**, with second-best results underlined. Original and gold model scores are highlighted for reference.

Method	de-De										fr-FR									
	XLS-R 128					XLS-R 53-DE					XLS-R 128					XLS-R 53-FR				
	$F1_T$	$F1_F$	MIA	GUM	Speedup	$F1_T$	$F1_F$	MIA	GUM	Speedup	$F1_T$	$F1_F$	MIA	GUM	Speedup	$F1_T$	$F1_F$	MIA	GUM	Speedup
Orig.	.584	.841	.621	.000	1.000×	.778	1.000	.622	.000	1.000×	.410	.572	.629	.000	1.000×	.756	1.000	.635	.000	1.000×
Gold	.566	.529	.513	.000	1.000×	.745	.706	.493	.000	1.000×	.469	.460	.509	.000	1.000×	.772	.800	.520	.000	1.000×
FT	.498	**.548**	.543	.588	34.34×	.661	.905	.585	.464	17.79×	.400	**.465**	**.539**	.545	18.12×	.759	.974	.627	.255	18.42×
NG	.550	.726	.562	**.797**	**1078**×	.764	.957	.587	**.643**	**558.7**×	.317	.349	.564	**.749**	**597.3**×	.768	**.935**	**.617**	**.501**	**610.2**×
NG+	.540	.567	**.487**	.522	16.89×	**.759**	**.878**	**.568**	.431	8.770×	.382	.008	.882	.000	8.900×	.759	.943	.620	.317	9.230×
CF-k	.587	.865	.622	.000	109.9×	.777	1.000	.616	.208	56.93×	**.436**	.594	.612	.414	58.23×	.770	1.000	.624	.338	58.86×
UNSIR	**.565**	.788	.616	.197	27.46×	.785	1.000	.619	.114	14.23×	.420	.591	.620	.259	14.67×	.768	1.000	.633	.089	14.94×
BT	.584	.789	.582	.489	20.02×	.726	.945	.585	.418	10.41×	.411	.583	.597	.409	10.60×	**.772**	.981	.621	.317	10.82×
BT-L	.584	.786	.576	.523	24.87×	.729	.948	.587	.434	12.94×	.412	.574	.591	.447	13.18×	.727	.981	.623	.306	13.42×
SCRUB	.584	.780	.600	.429	26.86×	.781	1.000	.615	.211	13.43×	.409	.532	.611	.358	13.68×	.769	1.000	.633	.089	13.94×

Negative Gradients (NG) consistently achieves the highest GUM scores across all configurations. For wav2vec 2.0, it surpasses the second-best method by 35% on FSC and 26% on SLURP*. On multilingual tasks with XLS-R 53, the improvements are even more evident: 39% for both ITALIC and German SpeechMASSIVE, and 48% for French SpeechMASSIVE. These gains stem from exceptional computational efficiency (up to 1748× speedup on FSC) combined with strong forgetting effectiveness, particularly on multilingual datasets.

NegGrad+ (NG+) occasionally achieves better task performance than NG, matching its MIA scores. However, its GUM scores suffer due to significantly lower computational efficiency. The method also exhibits catastrophic forgetting in some cases, particularly with XLS-R 128, where F1 scores drop hugely (0.001 on ITALIC, 0.008 on French SpeechMASSIVE).

Fine-Tuning (FT) shows particular strength with complex architectures. On ITALIC with XLS-R 128, it nearly matches the gold model's performance (F1 0.638 vs 0.643). However, its full-network updates limit efficiency, with speedups ranging

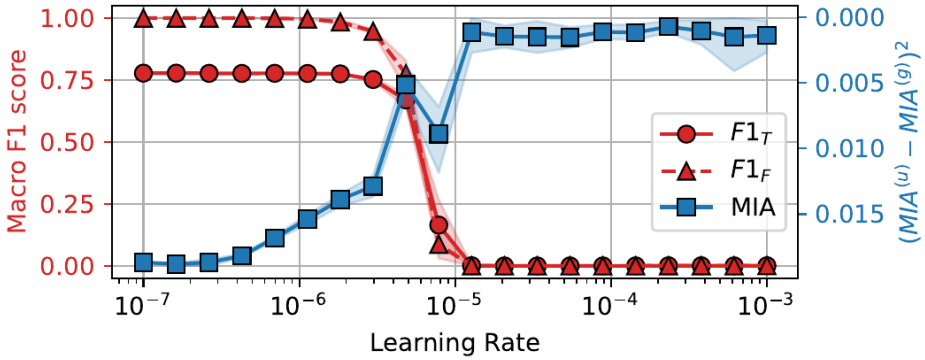


Fig. 4.6 **Trade-off between utility and efficacy.** Relationship between utility (test and forget F1) and efficacy (MIA) on NG as the learning rate varies (ITALIC, XLS-R 53-IT).

only from 8× to 84×.

Catastrophic Forgetting (CF-k) presents an interesting trade-off. While achieving the second-best efficiency scores, its layer-selective approach risks incomplete unlearning. This manifests in consistently higher MIA scores compared to gold models, particularly visible in SpeechMASSIVE (0.612-0.624 vs gold 0.493-0.520).

Bad Teaching variants show strong dataset dependence. They perform well on English datasets (FSC, SLURP) but struggle with multilingual models on ITALIC and SpeechMASSIVE.

Both SCRUB and UNSIR achieve poor GUM scores due to modest speedups (6-65×) and inconsistent forgetting effectiveness.

While recent literature [184, 185, 223, 224] has focused primarily on efficacy and utility, our results highlight the importance of efficiency through GUM's integrated evaluation. The consistent strong performance of NG across all metrics validates its position as a well-rounded unlearning approach.

Learning rate analysis. The learning rate significantly influences unlearning effectiveness under fixed computational budgets. Lower rates preserve model utility but limit forgetting effectiveness. Higher rates achieve better forgetting but risk degrading overall performance. Figure 4.6 demonstrates this utility-efficacy trade-off on ITALIC using NG across different learning rates.

GUM metric validation. Table 4.10 compares GUM against NoMUS [184], which averages model accuracy and MIA scores.

This comparison reveals several key advantages of GUM. First, GUM correctly assigns zero scores to trivial approaches (original and gold models) that NoMUS incorrectly rewards. Second, when UNSIR's efficacy deteriorates below the original model, GUM appropriately assigns a zero score, while NoMUS gives a misleadingly high score (0.700) by failing to contextualize MIA performance.

Table 4.10 **GUM vs. NoMUS comparison on SLURP*.** Comparison of MU techniques using GUM and NoMUS on the wav2vec 2.0.

Method	$F1_T$	MIA	Speedup	NoMUS	GUM
Orig.	.689	.628	1.000×	.717	.000
Gold	.707	.506	1.000×	.848	.000
NG	.695	.604	1748×	.744	.563
UNSIR	.673	.637	64.07×	.700	.000
SCRUB	.697	.608	64.82×	.741	.429

Third, GUM successfully differentiates methods with similar utility and efficacy based on efficiency. For instance, NG and SCRUB achieve similar NoMUS scores, but GUM (0.563 vs. 0.429) reflects NG's substantial efficiency advantage (1748× vs. 65× speedup).

Training duration impact. Table 4.11 examines how training duration affects the utility-efficacy trade-off using SLURP* with NG+. Extended training (60 epochs) achieves near-gold utility (F1 0.696 vs. 0.707) but compromises forgetting effectiveness. The high MIA score (0.611 vs. original 0.628) indicates persistent memorization of forgotten data, suggesting rigid decision boundaries that resist unlearning.

Table 4.11 **Effect of training duration on unlearning difficulty, wav2vec 2.0, SLURP*.** Each experiment uses NG+ with LR = 5e-07.

Epochs	$F1_T$	$F1_T^{(g)}$	MIA	$MIA^{(g)}$	$MIA^{(o)}$	GUM
5	.395	.398	.496	.510	.561	.678
7	.383	.419	.524	.515	.566	.680
11	.499	.487	.480	.492	.593	.686
15	.564	.550	.538	.491	.589	.644
60	.696	.707	.611	.506	.628	.421

Shorter training periods (5-15 epochs) better align with gold model behavior (MIA 0.480-0.538 vs. gold 0.491-0.515). We identify 11 epochs as an optimal point, balancing adequate utility (F1 = 0.499) with effective forgetting (MIA = 0.480) before overfitting sets in.

These findings demonstrate that successful unlearning requires carefully calibrated training duration to balance learning effectiveness against memorization permanence.

4.4.4 Summary and Practical Implications

UnSLU-BENCH provides the first unified evaluation framework for studying machine unlearning in spoken language understanding systems. Through a consistent application of the GUM metric, our experiments reveal three central insights. First, the trade-off between forgetting efficacy and utility is highly sensitive to optimization hyper-parameters, particularly the learning rate and training duration. Second, while several methods achieve comparable efficacy or utility in isolation, efficiency remains the decisive differentiator in real-world feasibility. Third, Negative Gradients (`NG`) consistently demonstrates balanced performance across all datasets and model architectures, achieving high GUM scores due to its strong forgetting capacity and exceptional computational speedups.

These findings have important implications for both researchers and practitioners. UnSLU-BENCH establishes a reproducible and interpretable standard for assessing unlearning methods beyond simple accuracy-privacy trade-offs. The GUM metric enables fine-grained comparisons across approaches and highlights inefficiencies that would otherwise remain hidden in average-based evaluations. For deployed speech systems, the benchmark underscores that effective unlearning does not require full retraining, as methods like `NG` offer meaningful privacy compliance with minimal computational overhead. Finally, the observed sensitivity to training dynamics suggests that unlearning should be treated not as a one-time correction, but as an iterative optimization process, integrated directly into model maintenance pipelines.

4.5 ARCH: A Unified Benchmark for Audio Representation Learning

Modern audio processing requires models that can understand diverse acoustic signals, from environmental sounds to music and speech. However, existing benchmarks often focus on single domains, making it difficult to assess models' general-purpose capabilities. We present ARCH (**A**udio **R**epresentation ben**CH**mark), a comprehensive benchmark designed to evaluate audio representations across multiple domains and tasks.

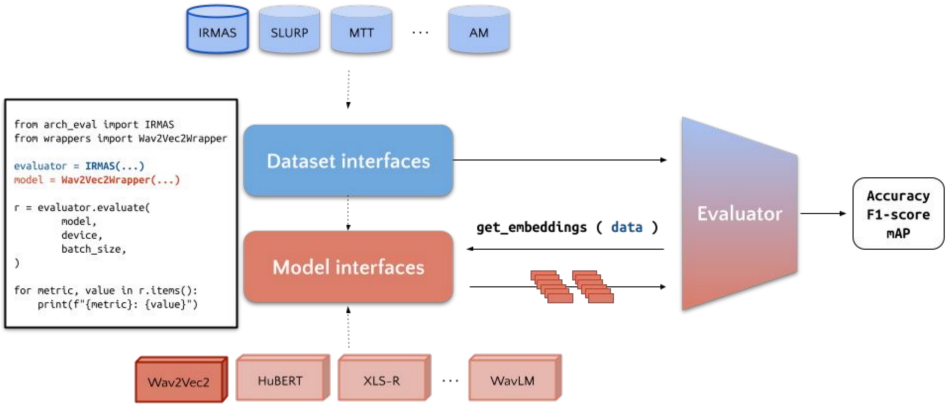


Fig. 4.7 **ARCH.** Overview of the proposed evaluation framework.

4.5.1 Framework Architecture

ARCH implements a modular design to facilitate expansion and customization while maintaining consistent evaluation protocols. The framework, built in Python, provides standardized interfaces for both dataset integration and model evaluation. An overview is depicted in Figure 4.7.

Dataset integration. Each dataset is encapsulated in a dedicated class that handles: (i) data loading and pre-processing; (ii) metadata management; (iii) batch iteration logic; and (iv) task-specific evaluation procedures.
This modular approach allows new datasets to be added by implementing a single class that conforms to the standard interface. The framework supports both single-label and multi-label classification tasks through standardized training and evaluation loops. Dataset-specific requirements, such as custom data splitting strategies, can be implemented while maintaining the common evaluation protocol.

Model integration. Models are integrated through wrapper classes that expose a consistent interface for embedding generation. Each wrapper must implement a method that converts raw audio input into sample-level embeddings. This abstraction separates model-specific processing from the general benchmarking workflow, enabling straightforward addition of new architectures.

Evaluation Methodology

Our evaluation protocol focuses on assessing the inherent quality of learned representations rather than optimizing task-specific performance. This philosophy guides several key methodological choices.

Representation processing. All models in our evaluation generate frame-level vector representations. We convert these sequential representations into fixed-dimension vectors through average pooling across frames. This standardization ensures fair comparison across different architectures.

Classification setup. We evaluate representations through a simple linear classification layer to assess their intrinsic discriminative power. The training runs for 200 epochs using AdamW optimizer, with the learning rate following a 10% warmup period reaching 0.001 and a linear decay applied for the remaining training period. Importantly, no model fine-tuning is permitted. The single linear layer classifier prevents complex nonlinear transformations from masking the quality of the base representations.

Design constraints. While the modular design of ARCH allows flexible model integration, we explicitly prohibit additional trainable parameters during embedding extraction, and complex pooling mechanisms (e.g., Attention Pooling). These restrictions may limit absolute performance but ensure fair comparison of fundamental representation quality. Our goal is to evaluate the inherent ability of the models to capture relevant acoustic information, providing objective guidance for model selection in specific applications.

Datasets Overview

ARCH incorporates twelve datasets across three audio domains, ensuring comprehensive evaluation of representation capabilities. Table 4.12 provides detailed dataset specifications.

Environmental Audio Collections. Our acoustic events evaluation uses four established datasets: ESC-50 [225] provides environmental sound classification; UrbanSound 8K (`US8K`) [226] focuses on urban acoustics; FreeSound Dataset 50K (`FSD50K`) [227] offers multi-label audio event detection; and VIVAE [228] addresses affective sound recognition.

Table 4.12 **ARCH dataset collection.** Overview of datasets included in the ARCH framework, showing their domains (acoustic events , music , speech), task types (single-label *S*, multi-label *M*), number of samples, average duration, and class count.

Dataset	Domain	Task	Samples	Avg duration	Classes
ESC-50 [225]		S	2000	5.0s	50
US8K [226]		S	8732	3.61s	10
FSD50K [227]		M	51197	7.64s	200
VIVAE [228]		S	1085	0.90s	6
FMA [229]		S	8000	29.98s	8
MTT [230]		M	21108	29.12s	50
IRMAS [231]		M	8278	5.73s	11
MS-DB [232]		S	21571	2.97s	8
RAVDESS [95]		S	1440	3.70s	8
AM [233]		S	30000	0.64s	10
SLURP [147]		S	72396	2.85s	77
EMOVO [234]		S	588	3.12s	7

These collections span diverse acoustic scenarios, with sample durations ranging from 1 to 7.5 seconds approximately. While FSD50K allows multiple labels per sample, the other three implement single-label classification.

Music Processing Tasks. The music domain evaluation employs four complementary datasets: Free Music Archive (`FMA`) [229] and MagnaTagATune (`MTT`) [230] target genre classification; IRMAS [231] focuses on instrument recognition; and Medley-solos-DB (`MS-DB`) [232] evaluates musical style detection.
This selection enables assessment across different musical analysis tasks. MTT and IRMAS support multi-label classification, while FMA and MS-DB use single labels. Sample durations show significant variation, from 3 to 30 seconds.

Speech Understanding. From the wide range of available speech datasets, we select four collections that reflect complementary challenges and characteristics: RAVDESS [95] and EMOVO [234] focus on emotion recognition; AudioMNIST (`AM`) [233] provides spoken digit classification; and SLURP [147] tests spoken language understanding through intent classification.
All speech datasets implement single-label classification, with samples ranging from 0.6 to 3.7 seconds.

We maintain original data splits where provided, otherwise implementing consistent cross-validation or fixed partitions across all evaluated models.

Models Overview

We evaluate five state-of-the-art self-supervised learning models: wav2vec 2.0 (`W2V2`) [1], WavLM [57], HuBERT [2], data2vec (`D2V`) [235][6], and XLS-R [14]. These models share several key architectural characteristics that make them particularly suitable for our evaluation. All models employ transformer-based architectures, leveraging the power of self-attention for audio processing. Each model undergoes pre-training on extensive speech datasets, though this speech-centric training may affect their generalization to other audio domains. The models operate directly on raw audio waveforms rather than preprocessed spectrograms or other intermediate representations. For feature extraction, they employ a CNN front-end that processes the raw waveform into frame-level features. These features then feed into a transformer encoder that captures contextual relationships across the audio sequence.

For each architecture, we evaluate three model scales: base (`B`), large (`L`), and extra-large (`XL`). These variants range from approximately 100M parameters in base models to 300M in large models and 1B in extra-large configurations. The models leverage diverse pre-training datasets including LibriSpeech [59], Libri-Light [236], GigaSpeech [205], and VoxPopuli [63].

To address a persistent limitation in the field, i.e., the lack of publicly available models pre-trained on diverse, non-speech audio, we additionally release a new suite of models of varying sizes pre-trained on the large-scale AudioSet [237] corpus. This effort directly addresses the current imbalance in audio research, where most existing pre-trained models are optimized for speech-related objectives and therefore transfer poorly to other acoustic domains such as environmental sound or music. By extending pre-training beyond linguistic signals, our models enable systematic investigation of general-purpose audio representation learning and cross-domain transferability.

[6] data2vec [235] introduces a self-supervised framework that unifies learning across modalities (speech, vision, and text) by predicting latent representations of the full input data. For speech, it extends masked prediction by using teacher-student training to learn contextualized representations of the entire signal rather than just masked regions.

Table 4.13 **Performance of SSL models on the ARCH framework.** Models marked with ◇ were pre-trained on AudioSet. The best overall scores are highlighted with a colored background, while the best results within each model size category are shown in **bold**.

Model	Size	Acoustic Events				Music				Speech			
		ESC-50	US8K	FSD50K	VIVAE	FMA	MTT	IRMAS	MS-DB	RAVDESS	AM	SLURP	EMOVO
W2V2	B	45.73	55.48	19.39	31.47	50.54	37.56	35.14	66.06	55.32	86.38	14.37	31.80
WavLM	B	49.88	61.84	17.63	36.31	48.71	34.93	32.62	54.18	**67.94**	99.50	30.98	**43.08**
WavLM+	B	58.73	64.07	21.57	36.17	56.17	38.24	35.76	57.51	52.20	**99.63**	28.06	36.73
HuBERT	B	58.90	67.28	24.53	**40.48**	54.63	38.78	36.65	58.46	65.28	99.58	33.75	40.48
D2V	B	23.63	45.63	10.06	30.19	40.58	27.60	25.87	50.74	48.03	99.06	**43.57**	27.27
◇W2V2-AS	B	52.61	70.48	21.29	31.26	59.50	37.92	35.85	64.61	45.94	88.09	11.00	30.83
◇HuBERT-AS	B	**68.80**	**79.09**	**31.05**	40.06	**65.87**	**43.44**	**47.67**	**67.81**	63.54	98.84	20.53	33.39
W2V2	L	13.13	42.70	5.80	22.01	41.71	20.95	19.91	50.23	11.57	45.74	7.33	19.27
XLS-R	L	51.28	69.96	23.71	36.28	56.96	38.28	38.42	66.71	31.48	98.88	12.74	20.35
WavLM	L	67.20	70.92	32.21	42.51	61.13	41.29	42.53	68.00	71.76	99.75	42.34	**45.29**
HuBERT	L	63.98	70.00	29.51	40.95	54.79	38.36	36.81	64.08	72.57	**99.95**	**45.26**	43.76
D2V	L	25.35	49.15	10.82	30.57	43.46	28.52	27.08	44.20	45.14	99.15	28.60	23.07
◇W2V2-AS	L	**74.39**	**79.00**	**37.58**	39.65	66.58	**44.51**	49.87	76.90	59.49	99.42	17.74	38.20
◇HuBERT-AS	L	71.52	75.63	37.41	**44.28**	**67.54**	43.35	**50.46**	**77.82**	**73.26**	99.59	20.46	38.61
XLS-R	XL	**66.95**	**75.90**	**31.61**	40.41	**62.79**	**41.99**	**43.57**	**69.79**	55.44	99.86	25.14	34.58
HuBERT	XL	63.40	69.66	29.32	**42.72**	56.25	37.76	37.30	64.71	**75.69**	**99.95**	**47.81**	**47.17**

While other transformer models operating on spectrograms exist [238, 239], we focus on waveform-based approaches to maintain consistent comparison. This constraint allows us to examine how speech-focused pre-training affects generalization to other audio domains. Our evaluation aims to guide domain-specific model selection and inform the development of more sophisticated architectures.

4.5.2 Results and Discussion

Our evaluation follows standard practices [199], using mean average precision (mAP) for multi-label tasks (FSD50K, MTT, IRMAS) and accuracy for single-label classification. Table 4.13 summarizes our findings across all domains.[7]

Acoustic Events Domain. Speech-pretrained models demonstrate surprisingly good generalization to environmental sounds. Among base-scale models, HuBERT and WavLM+ lead performance metrics, with HuBERT showing particular consistency. However, our AudioSet-trained variant (◇HuBERT-AS) significantly outperforms the speech-trained HuBERT on three of four datasets, highlighting the value of diverse pre-training data.

Scaling to larger models amplifies these benefits. The large wav2vec 2.0 model trained on AudioSet (◇W2V2-AS) achieves top performance on three datasets, closely

[7] github.com/MorenoLaQuatra/ARCH

followed by ◇HuBERT-AS. This demonstrates how increased model capacity combines with domain-appropriate pre-training data to enhance performance.
For extra-large models, we evaluate only speech-pretrained versions due to computational constraints. XLS-R shows modest gains over its large counterpart, likely due to its more diverse pre-training data. However, the performance of HuBERT plateaus at this scale, suggesting diminishing returns from pure scale without domain-appropriate data.

Music Domain. Our AudioSet-pretrained models dominate music-related tasks. Base ◇HuBERT-AS outperforms all other base models, while large ◇HuBERT-AS leads on three of four datasets. The exception is `MTT`, where ◇W2V2-AS shows slightly better results.
Model scaling consistently improves performance, particularly evident in instrument detection on `MS-DB`, where AudioSet pre-trained models show a 14% improvement from base to large sizes. However, even extra-large speech-pretrained models fail to match AudioSet-trained variants, emphasizing the crucial role of pre-training data diversity.
◇HuBERT-AS consistently outperforms ◇W2V2-AS at both base and large scales, suggesting that HuBERT's discrete target pre-training objective better captures musical features.

Speech Domain. Speech-pretrained models, unsurprisingly, excel on speech tasks, significantly outperforming AudioSet-trained variants. Model scaling provides substantial benefits, exemplified by an 8% accuracy gain on RAVDESS when moving from base to large models. The extra-large HuBERT achieves best-in-class performance across all speech datasets.
At base scale, WavLM leads performance metrics, leveraging its masked prediction pre-training. For large models, both WavLM and HuBERT show superior results, suggesting their discrete target approaches scale effectively.
Notably, on the Italian EMOVO dataset, English-pretrained HuBERT outperforms the multilingual XLS-R, indicating strong cross-lingual generalization capabilities.

4.5.3 Summary and Practical Implications

ARCH provides a unified and extensible framework for evaluating audio representation learning across multiple domains. By standardizing dataset integration and

evaluation procedures, it enables fair comparison between models trained on different types of audio data and isolates representational quality from task-specific optimizations.

Our results highlight that the diversity of pre-training data is a key determinant of cross-domain performance. Models trained solely on speech transfer poorly to environmental and musical sounds, while AudioSet-pretrained variants generalize more effectively beyond linguistic content. Scaling model size improves performance across all domains, though these gains diminish without sufficiently diverse data. HuBERT-based approaches, in particular, demonstrate strong and consistent generalization, validating their discrete target pre-training strategy.

These findings carry practical implications for both research and application. ARCH offers a principled foundation for selecting models suited to specific audio domains and provides an open, reproducible platform for studying general-purpose audio representations. By releasing all code, datasets, and AudioSet-pretrained models under open licenses, we aim to support transparent benchmarking and accelerate progress toward more universal, domain-agnostic audio understanding systems.

4.6 Conclusions

This chapter has presented three comprehensive frameworks for evaluating critical aspects of speech and audio foundation models. Each framework addresses a distinct challenge in developing robust, responsible, and broadly applicable AI systems.

SHALLOW (§4.3) represents the first systematic approach to quantifying and characterizing hallucinations in ASR systems. By decomposing errors into lexical, phonetic, morphological, and semantic dimensions, SHALLOW reveals nuanced failure patterns that traditional WER metrics miss entirely. Our evaluation demonstrates that even high-performing models can produce dangerous hallucinations, particularly in critical domains like healthcare transcription.

The framework multi-dimensional analysis has important implications for model development. Models with similar WER scores often show drastically different hallucination patterns, suggesting that architectural choices significantly influence the types of errors produced. This insight enables more targeted improvements in

model design, particularly for applications where certain types of errors are more costly than others.

UnSLU-BENCH (§4.4) addresses the growing need for verifiable data removal from trained models. Our evaluation framework reveals complex trade-offs between forgetting effectiveness, computational efficiency, and maintained model utility. The introduction of the GUM metric provides the first comprehensive way to evaluate these competing objectives simultaneously.
Our results demonstrate that simple approaches like Negative Gradients can be surprisingly effective when all three aspects are considered. However, the relationship between training duration and unlearning success suggests that privacy considerations should influence initial model training strategies. This finding has broad implications for how we approach model development in privacy-sensitive domains.

ARCH (§4.5) provides important insights into the generalization capabilities of modern audio foundation models. Our comprehensive evaluation across environmental sounds, music, and speech reveals both the promise and limitations of current approaches. While models pre-trained on speech can generalize reasonably well to other domains, pre-training data diversity proves more meaningful than model scale for cross-domain performance.
The superior performance of HuBERT-based architectures, particularly when trained on diverse data, suggests that discrete target pre-training objectives may better capture general audio characteristics. This finding could guide the development of more versatile foundation models for audio processing.

4.6.1 Future Research Directions

These frameworks open several promising research directions.

For hallucination detection, integrating SHALLOW's metrics into model training could lead to architectures that maintain high performance while minimizing dangerous fabrications. The framework could also be extended to evaluate hallucinations in multilingual settings, where cultural and linguistic factors may influence error patterns.

In machine unlearning, developing more efficient techniques that optimize the GUM metric could make privacy-preserving updates practical for larger models. Research

into the relationship between model architecture and unlearning effectiveness could inform the design of more privacy-friendly foundation models.

For audio representation learning, ARCH suggests exploring new pre-training objectives that combine the benefits of discrete targets with diverse audio domains. Investigating the relationship between pre-training data composition and cross-domain generalization could lead to more efficient training strategies.

Collectively, these frameworks provide essential tools for developing the next generation of audio AI systems, ones that are not just powerful, but also reliable, responsible, and broadly capable across diverse applications.

Chapter 5

Towards Natural Conversational AI: Foundation Models and Resources

5.1 Introduction

Human conversation is not just about the words we say: it is a dynamic mix of speech, emotion, and subtle cues that together bring meaning to our interactions. When we converse, we laugh, sigh, and express emotions through vocal bursts. We adjust our speaking style based on cultural and linguistic backgrounds, often drawing from rich regional language varieties. We maintain emotional coherence across multiple conversation turns. These nuanced aspects of human communication present significant challenges for current conversational AI systems.

Despite impressive advances in speech and language technology, today's conversational systems still fail to capture the full spectrum of human-like interaction. Current models excel at processing clean, scripted speech but struggle with spontaneous expressions like laughter, crying, or emotional outbursts. They typically handle only a single language or standardized dialect, struggling with regional varieties and limiting their ability to serve linguistically diverse populations. Most importantly, they lack the ability to maintain emotionally coherent, engaging conversations across multiple turns. Additionally, these models operate as black boxes, making decisions without providing clear explanations of how they process and interpret speech input. This opacity makes it difficult to trust their outputs or understand their limitations.

These challenges create a significant gap between human-human communication and human-machine interaction.

The development of more natural conversational AI systems faces four critical issues. First, existing foundation models focus primarily on speech recognition and generation, largely ignoring the crucial role of non-verbal vocalizations in human communication. These vocal expressions, from laughter to sighs, carry essential emotional and social information that current systems fail to process effectively. Second, most conversational AI research and resources concentrate on English, neglecting the rich linguistic diversity of other languages, their regional varieties, and their cultural nuances. Third, available dialogue datasets lack the emotional depth and conversational complexity needed to train systems capable of maintaining engaging, context-aware interactions. Fourth, current speech models provide no interpretable explanations for their decisions, making it difficult to understand and improve their behavior.

In this chapter, we present four major contributions to address these fundamental limitations. First, we introduce voc2vec [28], a novel foundation model specifically designed for understanding non-verbal aspects of human communication. Pre-trained using self-supervised learning on 125 hours of carefully curated non-verbal audio data, voc2vec creates universal representations of human vocalizations. Our model significantly advances the state-of-the-art in tasks ranging from emotion recognition to baby cry detection, providing an important building block for more empathetic conversational systems.

Second, we release DeepDialogue [29], a comprehensive dataset of $40,150$ multi-turn conversations that captures the complexity of natural human interaction. This dataset is unique in its incorporation of 20 distinct emotions across 41 conversational domains, with carefully designed emotional progressions that mirror natural human dialogue. Each conversation is accompanied by both text and emotionally-appropriate speech representations, enabling the development of multimodal conversational systems able to maintain coherent emotional states across multiple turns.

Third, we introduce ITALIC [30], the first large-scale spoken language understanding dataset for Italian, comprising over $16,500$ utterances from 70 speakers across different regions. This resource not only enables the development of conversational systems for Italian speakers but also captures regional linguistic variations, advancing our ability to create culturally aware dialogue systems. Building on this effort, we

further conduct an extensive investigation into the automatic identification of Italian language varieties directly from speech signals [31]. We demonstrate the feasibility of distinguishing between different Italian varieties while uncovering important insights about how linguistic features vary across regions.

Fourth, we develop novel explainability techniques for speech classification models, making their decision-making processes transparent and interpretable [32]. Our framework combines word-level audio segment attribution with paralinguistic feature analysis, providing comprehensive insights into how models process both linguistic and acoustic aspects of speech. This enables better understanding and validation of model behavior, crucial for developing trustworthy speech systems.

Our extensive empirical evaluation demonstrates the impact of these contributions across multiple dimensions of conversational AI. voc2vec achieves remarkable improvements over existing models across six different vocalization tasks. When integrated into dialogue systems, it can enable more accurate recognition of emotional states and non-verbal cues, fundamental for natural interaction. DeepDialogue allows to train systems that can maintain emotional consistency and natural conversation flow across extended interactions. ITALIC enables Italian conversational systems to achieve performance comparable to English ones while preserving important cultural and linguistic nuances. Our analysis of Italian language varieties demonstrates successful identification of regional speech patterns. Our explainability framework proves highly effective, with user studies showing strong preference for our explanations over baselines. These results demonstrate significant progress in making speech models both more capable and more transparent.

The practical implications of our work extend beyond academic benchmarks. By enabling better understanding of non-verbal cues, our models can help create more empathetic virtual assistants. The improved handling of emotional states and cultural variations makes conversational systems more accessible and engaging for diverse user groups. Our explainability techniques provide crucial transparency for deploying these systems in real-world applications.

Together, these contributions represent significant progress toward more natural, inclusive, and transparent conversational AI. By addressing fundamental gaps in non-verbal understanding, emotional dialogue modeling, multilingual support, and model interpretability, our work provides essential building blocks for the next generation of conversational systems. These advances bring us closer to the goal of creating AI

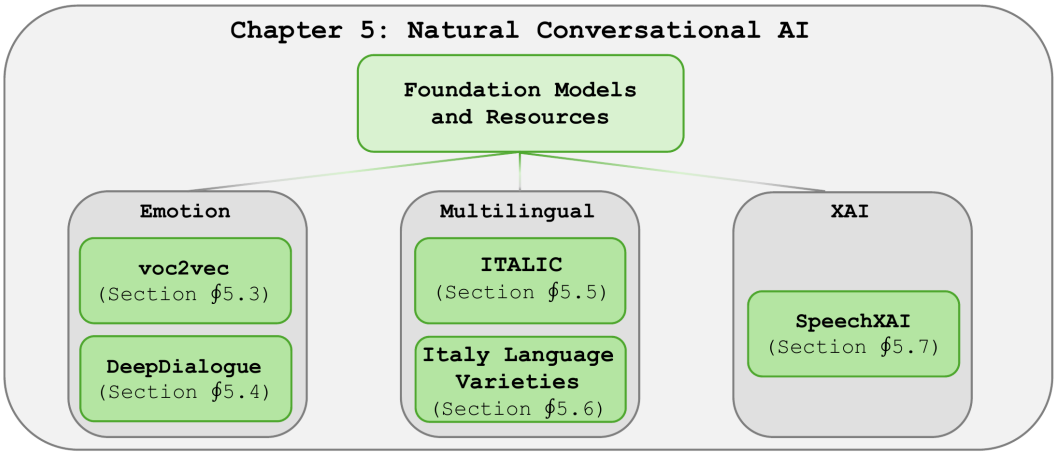


Fig. 5.1 **Chapter 5 Overview.** Graphical taxonomy of Chapter 5 topics.

systems that can engage in truly natural, emotionally aware, and culturally sensitive conversations while maintaining transparency in their decision-making processes.

The remainder of this chapter is organized as follows. Section §5.2 reviews relevant literature across speech processing, dialogue systems, and model interpretability. Section §5.3 introduces voc2vec, our foundation model for non-verbal vocalization understanding. Section §5.4 presents DeepDialogue, our large-scale emotional conversation dataset. Section §5.5 describes ITALIC, the first comprehensive Italian SLU dataset. Section §5.6 explores automatic identification of Italian language varieties. Section §5.7 presents our framework for explaining speech model decisions. Finally, Section §5.8 concludes the chapter and discusses future research directions. A graphical taxonomy of this chapter's contributions is provided in Figure 5.1.

5.2 Related Work

The development of natural conversational AI systems requires advances across multiple interconnected research areas. While recent years have seen remarkable progress in individual components, creating systems that can engage in truly natural dialogue, incorporating verbal and non-verbal communication, emotional awareness, and cultural sensitivity, remains a significant challenge. We review the key developments in each area and position our contributions within this scenario.

5.2.1 Foundation Models for Speech and Audio

The emergence of self-supervised foundation models has transformed speech processing capabilities. Models like wav2vec 2.0 [1] pioneered powerful contrastive learning approaches for speech representation learning. HuBERT [2] and WavLM [57] advanced this further through masked prediction objectives. While these models showed impressive performance on speech tasks, they remained primarily optimized for verbal content processing. The development of multilingual models like XLS-R [14] and Whisper [13] expanded cross-lingual capabilities but maintained focus on speech recognition rather than conversational dynamics.

In parallel, audio foundation models trained on AudioSet [237] demonstrated promise for general sound classification. However, these models struggle with the nuanced characteristics of human vocalizations, particularly non-verbal emotional expressions. This limitation highlights the need for specialized models that can capture the full spectrum of human communicative sounds.

5.2.2 Text-based Dialogue Systems

The evolution of dialogue systems has been primarily shaped by increasingly sophisticated datasets. Early efforts like KVRET [240] and M2M [241] focused on narrow domain-specific interactions between humans and machines. This scope expanded significantly with datasets like MultiWOZ [242], DailyDialog [243], and ABCD [244], which introduced multi-domain conversations that better reflect real-world scenarios.

Recent developments have focused on evaluating open-domain capabilities. Large-scale benchmarks like Chatbot Arena [245] and LMSYS-Chat-1M [246] enable systematic assessment of conversational agents. However, these text-only resources lack crucial elements of natural conversation including audio, emotion, and speaking style variations. This limitation has motivated research into more comprehensive multimodal dialogue systems.

5.2.3 Speech and Multimodal Dialogue

The incorporation of speech into dialogue systems has progressed through several generations of datasets and models. Early collections like DSTC2 [247] provided basic speech recognition annotations but were limited in scale and scope. DSTC10 [248] expanded this by including ASR hypotheses, while SpokenWOZ [249] introduced large-scale speech-text parallel data with comprehensive annotations.

More recent resources have focused on naturalistic conversation. DailyTalk [250] captures natural speech interactions, while MultiDialog [251] incorporates audio-visual content. However, these datasets often lack systematic coverage of emotional expression and speaking style variation, crucial aspects of human communication that extend beyond verbal content.

5.2.4 Emotional and Stylistic Expression

Understanding and generating appropriate emotional expressions is fundamental to natural conversation. Early research focused on acted emotional speech, with datasets like IEMOCAP [94], RAVDESS [95], and MELD [252] enabling basic research in emotional speech processing. These resources, while valuable, often relied on scripted performances that may not fully capture natural emotional dynamics.

Recent work has expanded both the scope and naturalism of emotional speech data. Expresso [253] and StyleTalk [254] have introduced more diverse speaking styles and expressive variations. Specialized evaluation frameworks like SD-Eval (Emo) [255] and E-Chat200 [256] support more nuanced assessment of emotional expression in dialogue. However, these advances primarily focus on verbal emotional expression, overlooking the crucial role of non-verbal vocalizations.

5.2.5 Non-verbal Vocalization Understanding

Non-verbal vocalizations represent a fundamental but understudied component of human interaction. Traditional speech processing research has focused primarily on verbal content and prosody [257, 258]. However, studies like [259, 260] have demonstrated that non-verbal cues may actually carry stronger emotional signals than speech itself.

Current models struggle with accurate classification of these sounds. Most work has focused on specific types of vocalizations, particularly laughter [261, 262]. While recent efforts like [263] have attempted broader coverage through large proprietary datasets, the lack of open resources has hindered progress. Some approaches like [264] have experimented to leverage speech SSL models for vocalization tasks, but require complex classifier chains to capture emotional dependencies.

5.2.6 Multilingual and Cross-cultural Conversation

The development of truly multilingual conversational AI faces unique challenges beyond simple translation. While datasets like MASSIVE [265] provide intent annotations across many languages, they often lack corresponding audio recordings. For Italian specifically, early efforts like AlmaWave-SLU [266] attempted machine translation approaches that failed to capture cultural nuances.

More recent work has begun addressing these limitations through language-specific resources. Models like BART-IT [267], IT5 [268], and Llamantino [269] have improved Italian language understanding capabilities. Studies of regional variation, exemplified by the DiatopIt corpus [270] and GeoLingIT task [271], have highlighted the importance of preserving linguistic diversity.

Specialized speech datasets like EMOVO [234] and IDEA [272] target specific aspects of Italian speech processing. However, even comprehensive collections like Common Voice [60] and Fleurs [273] lack the task-specific annotations needed for dialogue systems.

5.2.7 Model Interpretability and Responsible Development

As speech and dialogue systems become more sophisticated and widely deployed, understanding their decision-making processes becomes crucial. While previous sections have shown significant advances in model architectures and training data, these developments have made models increasingly complex and opaque. This opacity is particularly problematic for speech processing, where models must integrate multiple information streams including verbal content, prosody, and non-verbal cues discussed earlier.

Early attempts at speech model interpretation focused on visualizing spectrograms [274, 275]. While these approaches helped experts analyze frequency patterns, they proved difficult for non-specialists to understand. This limitation is particularly relevant for dialogue systems, where interpretability is essential for both developers and end-users. Recent work has explored more intuitive explanation methods aligned with how humans process speech. Some approaches focus on temporal explanations, either at the sample level [274] or using fixed-width segments [276]. SoundLIME [277] adapted LIME [278] for audio by segmenting in time or frequency domains. However, these methods often lack semantic grounding in spoken words, making them less useful for understanding higher-level speech understanding tasks.

More recent efforts have attempted to bridge this gap between signal processing and semantic understanding. Work on phoneme-level explanations [279] provides finer granularity but requires specialized annotations.

Beyond interpretability, the responsible development of conversational AI requires addressing broader ethical concerns. The multilingual challenges discussed earlier intersect with fairness issues, as studies have identified systematic biases based on accent [109], age [65], and other factors [3]. Work on bias detection [16] and mitigation [4, 22, 108] has proposed various solutions. However, balancing model transparency, fairness, and performance remains an ongoing challenge, especially for the complex foundation models and dialogue systems described above.

5.2.8 Positioning Our Contributions

Our work advances conversational AI through several complementary contributions that address critical gaps in existing research.

First, voc2vec provides the first foundation model specifically designed for non-verbal vocalization understanding. Unlike previous approaches that struggled with non-verbal sounds, our model learns rich representations of these communicative signals through self-supervised training on diverse open-source data.

Second, DeepDialogue offers unprecedented resources for modeling natural conversation flow. Its careful design of emotional trajectories across several domains enables more nuanced interaction modeling. The parallel text and emotionally-appropriate speech representations support development of truly multimodal systems that can maintain coherent emotional states.

Third, ITALIC enables robust Italian conversational AI while preserving important regional linguistic variations. Unlike previous resources that treated Italian as monolithic, our dataset captures regional speech patterns essential for culturally-aware systems. Building on this foundation, our detailed analysis of Italian language varieties demonstrates effective identification of regional speech patterns through innovative applications of contrastive learning.

Finally, our explainability framework provides interpretable insights into speech model behavior through both word-level audio segments and paralinguistic features. This enables better understanding of how models process both verbal content and acoustic characteristics, essential for developing more transparent systems.

Together, these advances represent significant progress toward conversational AI systems that can engage in natural, emotionally appropriate dialogue across languages and cultures, while remaining interpretable and trustworthy.

5.3 voc2vec: A Foundation Model for Non-Verbal Vocalizations

Understanding vocal emotional behavior is essential for conversational technologies, such as digital assistants and therapeutic tools, that aim to interact naturally with humans and anticipate their needs [263, 280, 281]. While speech prosody works in tandem with words to communicate emotions, vocal bursts are standalone sounds that express emotion without speech. Despite their significance, current models struggle to accurately classify vocal bursts, often focusing only on laughter and overlooking other meaningful vocalizations like sighs, gasps, or different types of laughs. In response to these limitations, we introduce voc2vec [28], a novel foundation model tailored specifically for non-verbal audio processing. Pre-trained using SSL on a diverse set of 10 open-source non-verbal audio datasets totaling 125 hours, voc2vec is designed to capture the unique features of non-verbal sounds. In the context of this work, the term foundation model refers to the model's architectural capability to learn robust, task-agnostic representations through self-supervision. This definition highlights the model's generalizability and feature-extraction power across multiple downstream tasks, rather than the massive volume of training data typically associated with text-based Large Language Models.

5.3.1 Methodology

Model Architecture

Voc2vec builds on the proven wav2vec 2.0 architecture [1], adapting it specifically for non-verbal sound processing. At its core, the model employs a dual-stage architecture: a CNN encoder for feature extraction and a transformer network for contextual modeling.

The CNN encoder processes raw audio waveforms into compact representations. Given an input sequence $X = \{x_1, x_2, \ldots, x_T\}$, it produces latent vectors:

$$\mathbf{Z} = \text{Encoder}(X) = \{z_1, z_2, \ldots, z'_T\}, \quad T' < T \tag{5.1}$$

where each $z_t \in \mathbb{R}^d$ encodes frame-level acoustic features and T' represents the compressed temporal dimension.

The transformer network then processes these encodings to capture long-range dependencies, outputting contextualized representations $\mathbf{C} = \{c_1, c_2, \ldots, c'_T\}$. This processing allows voc2vec to model both fine-grained acoustic patterns and broader temporal structures in non-verbal sounds.

Pre-training Strategy

The key innovation of voc2vec lies in its pre-training approach, which focuses exclusively on non-verbal vocalizations. We curated a diverse collection of 10 open-source datasets, totaling 125 hours of audio, that span the full spectrum of human non-verbal expression. While smaller than typical speech corpora, this collection provides focused exposure to the specific patterns and variations of non-verbal sounds. A summary of such datasets statistics is provided in Table 5.1.

AudioSet (vocalization split) [237] represents the broadest collection of non-verbal sounds in natural settings. Its carefully selected subset includes spontaneous human vocalizations like laughter, crying, and other vocal expressions, providing essential real-world variability.

FreeSound (babies split) [282] provides an extensive collection focusing on infant vocal expressions. The dataset captures the full range of baby vocalizations, from

Table 5.1 **voc2vec, pre-training datasets.** Summary statistics of the pre-training datasets, including total duration (in hours), number of samples, and average sample length.

Dataset	Dur. (h)	# Samples	Avg Dur. (s)
AudioSet (vocalization) [237]	36.94	13439	9.90
FreeSound (babies) [282]	23.42	1450	58.15
HumanVoiceDataset	0.06	179	1.21
NNIME [283]	3.55	5596	2.28
NonSpeech7K [284]	6.72	6983	3.46
ReCANVo [285]	2.46	7077	1.25
SingingDatabase [286]	3.97	113	126.48
TUT (babies) [287]	13.17	1540	30.79
VocalSketch [288]	10.53	10705	3.54
VocalSound [289]	24.37	20985	4.18
Voc125 (Total)	**125.19**	**68067**	**6.67**

crying to laughter and babbling, making it particularly valuable for pediatric applications.

Human Voice Dataset[1] consists of specialized recordings exploring vocal control techniques. While compact in size, it offers unique insights into controlled vocal production, including detailed pitch variations and phonetic articulations.

NNIME [283] captures emotional expressions in natural conversations between Mandarin speakers. Its recordings of spontaneous interactions showcase how emotions manifest through non-verbal cues across different affective states, from joy to frustration.

NonSpeech7K [284] encompasses everyday human vocal sounds beyond speech. By including common sounds like breathing, coughing, and laughing, it helps the model recognize frequent non-verbal expressions encountered in daily interactions.

ReCANVo [285] features over 7,000 vocalizations from individuals with limited speech capabilities. These recordings provide insights into how emotions and needs are communicated through non-verbal channels when speech is not the primary mode of expression.

Singing Database [286] contains professional vocal performances from Chinese Opera traditions. These recordings showcase structured non-verbal expressions

[1] https://github.com/vocobox/human-voice-dataset

within a formal performance context, offering examples of highly controlled vocal techniques.

TUT [287] specializes in infant cry recordings under realistic conditions. Its distinctive feature is the inclusion of background noise, making it valuable for developing robust detection capabilities in real-world environments.

VocalSketch [288] presents a collection of human vocal imitations. Participants used their voices to recreate environmental sounds, providing unique examples of how humans adapt their vocal apparatus to mimic non-human sounds.

VocalSound [289] contributes over $21,000$ crowdsourced recordings of spontaneous expressions. This large-scale collection of natural reactions like laughter, sighs, and coughs helps ensure robust recognition of common non-verbal cues.

This diverse dataset collection ensures voc2vec develops robust representations across the full range of human non-verbal expression, from spontaneous emotional outbursts to controlled vocal productions. While the total duration is smaller than typical speech pre-training corpora, the focused nature of these datasets enables effective learning of non-verbal patterns.

5.3.2 Experimental Setup

Pre-training procedure. Voc2vec adopts and adapts the wav2vec 2.0 pre-training architecture for non-verbal audio processing. The feature encoder implements seven convolutional blocks, each utilizing 512 channels. These blocks employ carefully chosen strides $(5,2,2,2,2,2,2)$ and kernel widths $(10,3,3,3,3,2,2)$. This configuration results in an encoder output frequency of $49Hz$, with approximately $20ms$ between samples and a $25ms$ audio receptive field. Following wav2vec 2.0's base configuration, we implement 12 transformer blocks with a model dimension of 768, inner dimension of $3,072$, and 8 attention heads. The pre-training process runs on two A6000 GPUs over 10.6 days, completing $400k$ updates. All audio samples are standardized to 10-second lengths during pre-processing to ensure consistent model input. The final model checkpoint is selected based on the lowest validation loss achieved during training.[2]

[2] github.com/koudounasalkis/voc2vec

Table 5.2 **voc2vec, fine-tuning datasets.** Summary statistics of the fine-tuning datasets, including number of classes, total duration (in hours), number of samples, and average sample length.

Dataset	# Classes	Dur. (h)	# Samples	# Avg Dur. (s)
ASVP-ESD [290]	13	15.07	12625	4.30
ASVP-ESD (babies) [290]	7	2.91	1339	8.22
CNVVE [291]	6	0.2	921	0.78
Donate A Cry	5	0.88	457	6.93
NonVerbal Vocalization	16	0.6	800	3.10
VIVAE [228]	6	0.27	1085	0.90

To investigate the impact of different initialization strategies, we explore three distinct pre-training approaches: (i) *Scratch Training*, where we pre-train exclusively on our non-verbal vocalization datasets (`Voc125`), (ii) *Speech Transfer*, where we continue from a LibriSpeech-pretrained model [1], and (iii) *Audio Transfer*, where we build upon an AudioSet-pretrained model [27]. This comparison helps us understand whether focused non-verbal data can enhance models previously trained on larger speech or general audio datasets.

Fine-tuning protocol. For downstream task adaptation, we augment voc2vec with a randomly initialized output layer on top of the transformer. The fine-tuning protocol maintains consistent hyper-parameters across all experiments, with detailed information on our project repository. As the evaluation datasets come with no predefined splits, we apply a 10-fold cross-validation for robust performance assessment.

Evaluation datasets. We evaluate voc2vec across six diverse classification tasks, summarized in Table 5.2. ASVP-ESD [290] provides 12,625 real-world emotional vocalizations, sourced from movies, YouTube, and other media. It also contains 1,339 baby vocalizations (tested separately as ASPV-ESD babies). Donate a Cry[3] focuses on infant vocalizations, with baby cry recordings and both action-based (e.g., burping) and emotional states (e.g., discomfort) annotations. CNVVE [291] examines vocal expression diversity, containing 921 samples across 6 categories, sourced from 42 participants including both healthy and dysarthric speakers. It is important to assess assistive technology applications. NonVerbal Vocalization Dataset[4] provides general non-verbal classification, with diverse sound categories including laughing and crying, making it ideal for testing generalization across different vocalization

[3] https://github.com/gveres/donateacry-corpus
[4] https://www.openslr.org/99/

Table 5.3 **Impact of different initialization strategies for voc2vec.** Training from scratch (Voc125), speech transfer from LibriSpeech (LS960) or audio transfer from AudioSet (AS). Reported are mean ± standard deviation for UAR, Accuracy, and Macro F1 across six downstream datasets. For reference, results for wav2vec 2.0 and HuBERT models pre-trained on LibriSpeech are also included. Best results in **bold**.

Model	Pre-training DS	UAR	Accuracy	F1 Macro
`voc2vec`	Voc125	.612±.212	.729±.146	.580±.230
`voc2vec-as`	AS+Voc125	.603±.183	.754±.131	.574±.194
`voc2vec-ls`	LS960+Voc125	**.661±.206**	**.802±.139**	**.636±.223**
`wav2vec2-ls`	LS960	.599±.237	.739±.192	.569±.259
`hubert-ls`	LS960	.627±.214	.784±.149	.611±.222

types. VIVAE [228] captures emotional intensity, with 1,085 human vocalizations from 6 distinct emotional classes. It offers a balanced representation of positive and negative states, with various intensity levels within each emotion.

Baselines and metrics. Our comparison includes several strong baselines derived from state-of-the-art architectures. For wav2vec 2.0, we test `wav2vec2-ls` pre-trained on LibriSpeech and `wav2vec2-as` pre-trained on AudioSet. Similarly, we evaluate HuBERT variants `hubert-ls` (LibriSpeech) and `hubert-as` (AudioSet). We also include WavLM [57], WavLM-plus, and data2vec [235], all in their base configurations with approximately 90M parameters. Additionally, we compare against two feature-based approaches: OpenSmile [258] and emotion2vec [292], both evaluated using two linear layers with ReLU activation. Performance is assessed through three complementary metrics. Unweighted Average Recall (UAR) measures the average recall across all classes, giving equal weight regardless of frequency. Accuracy provides the ratio of correct predictions to total instances. F1 Macro Score computes the harmonic mean of precision and recall, calculated per class and then averaged.

5.3.3 Results and Discussion

Impact of Pre-Training Strategies

Table 5.3 compares our three pre-training strategies: `voc2vec` (trained *from scratch*), `voc2vec-as` (*audio transfer* from AudioSet initialization), and `voc2vec-ls` (*speech*

transfer from LibriSpeech initialization). The results clearly establish `voc2vec-ls` as the superior variant across all metrics, indicating that speech-based pre-training provides an optimal foundation for non-verbal tasks[5]. Training from scratch (`voc2vec`) shows limited performance, highlighting the challenges of learning from our relatively small vocalization corpus alone. Interestingly, initializing from AudioSet (`voc2vec-as`) offers only marginal improvements over the scratch training, suggesting that general audio pre-training is less beneficial than speech-specific pre-training for vocal tasks.

When compared to models pre-trained solely on LibriSpeech, `voc2vec-ls` demonstrates substantial gains. It surpasses `wav2vec2-ls` by 10.3% in UAR and 11.7% in F1 Macro, underscoring the value of our focused vocalization pre-training even after speech pre-training. This improvement suggests that while speech pre-training provides useful acoustic foundations, explicit training on non-verbal sounds is crucial for capturing their unique characteristics.

Benchmark Performance

As shown in Table 5.4, `voc2vec-ls` achieves state-of-the-art performance across most datasets and metrics. We analyze the results across different vocalization types.

General non-verbal recognition. On the NonVerbal Vocalization dataset, `voc2vec-ls` reaches 0.872 UAR, outperforming the next-best model (`wavlm`) by 3.8%. This strong performance on general non-verbal classification demonstrates the model's ability to distinguish between diverse vocal expressions like laughing, crying, and other common sounds.

Emotional vocalization. For VIVAE, `voc2vec-ls` maintains clear superiority with 0.573 UAR and 0.558 F1 Macro, surpassing the strong `hubert-ls` baseline with a relative 9% improvement. The performance gap is particularly notable for subtle emotional expressions, where the model shows improved sensitivity to variations in emotional intensity.

[5] Subsequent to this work, we developed an additional voc2vec variant implementing the HuBERT pre-training approach. This version, also leveraging speech transfer, demonstrates a 6.6% relative improvement over our best wav2vec2-based voc2vec-ls model, resulting in an 11% improvement compared to the best hubert-ls baseline across all evaluation datasets. Complete results for this new variant are available in our project repository: https://github.com/koudounasalkis/voc2vec

Table 5.4 **voc2vec, fine-tuning results.** 10-fold cross-validation fine-tuning results across the six evaluated datasets. Best results in **bold**.

Model	UAR	Accuracy	F1 Macro	UAR	Accuracy	F1 Macro	UAR	Accuracy	F1 Macro
	ASPV-ESD			ASPV-ESD (babies)			CNVVE		
wav2vec2-as	.590±.016	.624±.014	.577±.016	.521±.126	.890±.044	.460±.132	.839±.060	.838±.063	.809±.073
wav2vec2-ls	.626±.014	.672±.013	**.627±.013**	.432±.171	.891±.059	.378±.140	.971±.018	.970±.017	.973±.016
hubert-as	.587±.173	.621±.194	.577±.183	.456±.121	.856±.055	.389±.106	.919±.058	.922±.058	.918±.059
hubert-ls	.622±.017	.664±.011	.619±.014	.515±.134	.924±.023	**.505±.141**	.972±.017	.972±.017	.972±.017
wavlm	.615±.016	.663±.012	.609±.016	.501±.151	.870±.063	.434±.144	.976±.018	.975±.019	.975±.019
wavlm-plus	.594±.017	.649±.009	.591±.016	.441±.133	.860±.080	.385±.155	.978±.015	.978±.015	.978±.015
data2vec	.585±.029	.606±.027	.567±.030	.514±.171	.880±.043	.468±.165	.981±.008	.980±.007	.981±.008
emotion2vec	.575±.042	.618±.012	.537±.018	.509±.206	**.926±.330**	.467±.175	.955±.022	.950±.023	.951±.023
opensmile	.106±.055	.125±.037	.040±.011	.281±.137	.785±.192	.258±.126	.206±.092	.249±.053	.160±.050
voc2vec-ls	**.633±.009**	**.673±.009**	**.627±.006**	**.526±.189**	.914±.039	.491±.175	**.982±.011**	**.982±.010**	**.982±.010**
Model	**NonVerbal Vocalization**			**Donate a Cry**			**VIVAE**		
wav2vec2-as	.619±.065	.607±.070	.575±.064	.282±.086	.783±.065	.254±.078	.449±.058	.448±.059	.412±.059
wav2vec2-ls	.838±.067	.817±.088	.818±.079	.360±.071	.709±.153	.295±.075	.365±.162	.375±.158	.320±.194
hubert-as	.837±.048	.826±.051	.803±.048	.364±.075	.743±.111	.304±.064	.455±.059	.464±.055	.415±.065
hubert-ls	.812±.047	.810±.047	.789±.054	.312±.067	.802±.051	.276±.062	.527±.067	.534±.062	.510±.071
wavlm	.840±.050	.832±.041	.820±.054	.352±.093	.721±.101	.307±.058	.230±.061	.256±.074	.161±.067
wavlm-plus	.731±.035	.728±.046	.676±.043	.349±.096	.835±.041	.306±.073	.194±.056	.211±.085	.107±.073
data2vec	.707±.076	.706±.078	.651±.073	.306±.106	.768±.100	.252±.081	.301±.096	.307±.101	.224±.117
emotion2vec	.764±.038	.760±.035	.726±.046	.210±.051	**.837±.036**	.228±.054	.513±.082	.516±.076	.502±.075
opensmile	.021±.007	.094±.030	.026±.015	.212±.050	.784±.155	.222±.062	.128±.077	.128±.025	.092±.025
voc2vec-ls	**.872±.049**	**.869±.036**	**.848±.049**	**.378±.173**	.793±.095	**.311±.092**	**.573±.062**	**.578±.055**	**.558±.060**

Infant vocalization. On ASVP-ESD baby subset, `voc2vec-ls` achieves significant improvements over existing approaches on UAR. While `emotion2vec` shows competitive performance on Donate a Cry (0.837 Accuracy), `voc2vec-ls` demonstrates more balanced performance across all metrics, suggesting better generalization to different types of infant vocalizations.

Clinical applications. In the CNVVE dataset, which includes both healthy and dysarthric speakers, `voc2vec-ls` shows robust performance across different speaker conditions. This suggests potential applications in healthcare and assistive technology contexts, where reliable recognition of non-verbal cues is crucial.

Comparison with existing approaches. When compared to other self-supervised approaches, `wav2vec2-ls` and `hubert-ls` show strong baseline performance but fall short on non-verbal tasks. WavLM and WavLM-plus, despite their sophisticated pre-training, achieve lower performance on most datasets, and data2vec, while effective for speech, shows limitations in capturing non-verbal patterns.
Traditional approaches show varying effectiveness. OpenSmile performs consistently below SSL-based models across all datasets, while emotion2vec, despite its special-

ization in emotion recognition, only outperforms `voc2vec-ls` in specific scenarios. The gap between feature-based and SSL approaches highlights the advantage of learned representations.

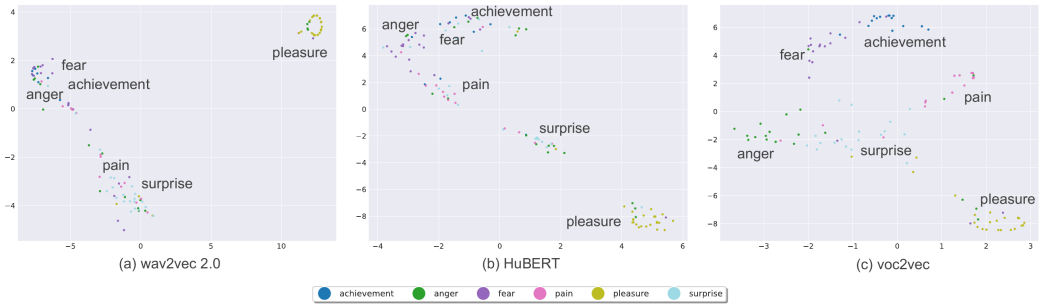


Fig. 5.2 **voc2vec, t-SNE visualization.** VIVAE dataset, first fold. Representations from `wav2vec-ls` (left), `hubert-ls` (center), and `voc2vec-ls` (right).

Representation analysis. Figure 5.2 visualizes t-SNE projections of VIVAE dataset embeddings from `wav2vec2-ls` (left), `hubert-ls` (center), and `voc2vec-ls` (right). The voc2vec representations show notably improved structure, with better intra-cluster cohesion and inter-cluster separation compared to both wav2vec 2.0 and HuBERT embeddings. This improved clustering suggests that voc2vec learns more discriminative features for non-verbal sounds.

5.3.4 Summary and Practical Implications

The strong performance of voc2vec across diverse tasks opens up numerous practical applications in real-world scenarios. In conversational AI systems, the model enables more nuanced emotion recognition, potentially leading to more natural and empathetic interactions. For pediatric applications, it offers more reliable infant cry detection and classification, supporting both parents and healthcare providers. The model's robust performance also makes it valuable for developing enhanced accessibility tools for individuals with speech impairments. In clinical settings, it can provide better understanding of non-verbal cues, supporting both diagnosis and patient monitoring.

These results demonstrate that focused pre-training on non-verbal vocalizations, combined with the strong foundation of speech pre-training, can significantly advance the state of non-verbal sound processing. Our comprehensive evaluation across

multiple tasks and datasets validates the effectiveness of our approach, with consistent improvements over existing methods regardless of the specific vocalization type or application context. The versatility of the model across different applications underscores the value of our targeted pre-training approach.

While these advances are significant, important challenges remain. The current implementation requires high-quality audio input and may struggle with extreme background noise or overlapping vocalizations. Future work could explore more robust architectures and data augmentation strategies to address these limitations. Nevertheless, voc2vec represents a significant step toward more comprehensive understanding of human vocal expression, providing a foundation for developing more naturalistic and emotionally aware AI systems.

5.4 DeepDialogue: A Large-Scale Emotional Conversation Dataset

While voc2vec advances our ability to process non-verbal emotional expressions, developing truly natural conversational AI systems requires understanding how emotions evolve throughout extended dialogues. Current dialogue datasets are predominantly text-only, lack emotional depth, and rarely capture the natural progression of feelings across multiple conversation turns. Even state-of-the-art language models, while impressive in single-turn responses, struggle to maintain emotional coherence in longer interactions.

To address these limitations, we present DeepDialogue [29], a multimodal dataset containing $40,150$ high-quality multi-turn dialogues. Each conversation spans one of 41 diverse domains and incorporates carefully designed emotional progressions across 20 distinct emotional states. Unlike existing resources, DeepDialogue provides both text and emotionally-appropriate speech representations, creating the first large-scale open-source dataset that preserves emotional context across extended conversations. By pairing different language models (4B-72B parameters) to generate initial dialogues and applying rigorous quality filtering, we create a resource that advances our understanding of both emotional dialogue generation and spoken conversation modeling.

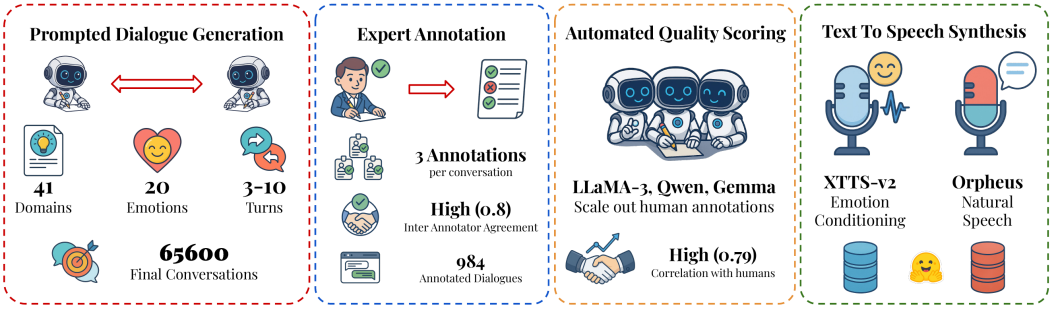


Fig. 5.3 **DeepDialogue dataset generation framework.** Overview of the data creation pipeline comprising four main stages: (1) dialogue generation conditioned on emotion and domain context, (2) human annotation for quality and consistency, (3) automated filtering using large language models, and (4) dual text-to-speech synthesis for diverse acoustic realizations.

5.4.1 Methodology

Our approach to creating DeepDialogue follows a systematic four-stage pipeline: domain and emotion design, dialogue generation with emotional progression modeling, quality evaluation and filtering, and speech synthesis. Each stage is carefully designed to ensure the creation of natural, emotionally coherent conversations across multiple modalities. Figure 5.3 shows the dataset generation framework.

Domain and Emotion Setup

We establish 41 distinct conversation domains that span a broad spectrum from concrete to abstract topics. Concrete domains like *travel*, *cars*, and *cooking* provide clear reference points for discussion, while abstract domains such as *philosophy*, *spirituality*, and *politics* challenge models with conceptual reasoning. This balanced domain selection covers everyday topics (*weather*, *work*), specialized knowledge areas (*science*, *finance*), and social-emotional contexts (*relationships*, *health*). The full list of domains can be seen in Figure 5.4.

The emotional framework encompasses 20 distinct emotions, carefully selected to represent the full range of human expression. These include basic emotions (*happiness, sadness, anger*), complex social emotions (*pride, embarrassment, gratitude*), and epistemic emotions (*curiosity, confusion, surprise*). We map emotions

Fig. 5.4 **DeepDialogue, domains.** Overview of domain distribution within the dataset.

to domains through a pilot study where human annotators rate the naturalness of expressing specific emotions within each domain context.

A key innovation is our structured mapping of emotion transitions, representing plausible progressions of emotional states across dialogue turns. These mappings ensure conversations evolve with emotionally realistic dynamics rather than exhibiting implausible emotional shifts. For example, transitions from *frustration* might lead to *anger*, *disappointment*, or *anxiety*, but would rarely jump directly to *excitement* or *happiness* without intermediate states. We developed a directed graph of permissible emotion transitions based on psychological literature on emotional progression [293, 294]. This approach allows us to construct "emotional arcs" for dialogues of different lengths, incorporating both predictable patterns (e.g., *curiosity* $\rightarrow$ *surprise* $\rightarrow$ *excitement*) and more complex trajectories that reflect the nuanced nature of human emotional expression. The resulting emotion transition framework provides a structured yet flexible mechanism for generating dialogues with coherent emotional progression while maintaining diversity in conversational dynamics. A graph depicting this transition framework is shown in Figure 5.5.

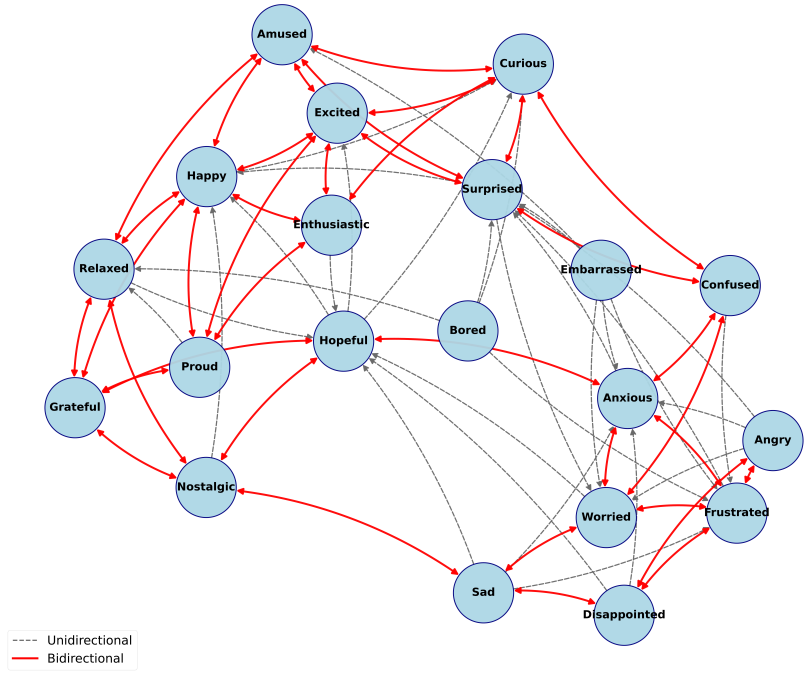


Fig. 5.5 **DeepDialogue, emotions.** Emotion transition graph in DeepDialogue.

Dialogue Generation Process

The generation process begins by randomly selecting a domain and initial emotion, using stratified sampling to ensure balanced representation. We then initiate conversations between pairs of language models selected from a pool of 9 instruction-tuned LLMs ranging from 4B to 72B parameters, creating 14 different possible pairings.[6]

Each turn is generated through carefully curated prompts that incorporate the complete conversation history, specific domain context, and target emotion for the current turn. We provide few-shot examples demonstrating appropriate emotional expression within similar contexts. To maintain natural conversational flow, we limit responses to 25 words maximum. The emotional progression follows our predefined transition rules while remaining sensitive to conversational context. This dynamic emotion selection process is guided by our emotion graph, which encodes allowed transitions and their associated probabilities, prioritizing contextually coherent emotional progressions. For example, in a *travel* domain dialogue where one speaker expresses

[6]We employ the following models: Llama-3.1-8B-Instruct [295], Llama-3.3-70B-Instruct [295], Qwen2.5-32B-Instruct [296], Qwen2.5-72B-Instruct [296], phi4-mini-instruct [102], phi-4 [297], c4ai-command-r7b [298], gemma3-4B-Instruct [299], gemma3-27B-Instruct [299].

excitement about an upcoming trip, the system might sample the next emotion from likely options such as *enthusiasm*, *curiosity*, or *amusement*, mirroring the natural flow of human emotional responses. This probabilistic strategy introduces natural variation while preserving emotional consistency across the dialogue. To further enhance contextual alignment, the system incorporates content-aware emotional routing based on the textual analysis of preceding turns. A lightweight sentiment analyzer [300] detects prominent emotional signals, prompting real-time adjustments to transition probabilities. For instance, if a message conveys personal achievement, the model increases the likelihood of selecting emotions such as *pride*, *curiosity*, or *excitement*, while reducing the probability of less congruent responses like *boredom* or *frustration*.

We introduce additional mechanisms that adjust emotional transition probabilities according to conversation length. Shorter dialogues (3-4 turns) tend to follow more focused emotional trajectories, while longer ones (7-10 turns) allow for richer affective evolution with smoother, more gradual shifts between emotions. To reflect natural conversational rhythms, each dialogue is assigned a random length within this range. All generated exchanges are designed to sound natural and engaging, more like friendly human conversations than system-driven interactions. To ensure this, the models are prompted to adopt a conversational persona and convey emotions not only through word choice but also through expressive linguistic cues such as exclamation marks for enthusiasm, ellipses for hesitation, and descriptive phrasing for emotional nuance. Example prompts are provided in the original paper [29].

This procedure is repeated across all domains, producing $1,600$ dialogues per domain, for a total of $65,600$ raw samples. The resulting dataset captures a wide range of conversational dynamics, systematically varying by domain, emotional flow, model pairing, and dialogue length, offering a rich foundation for studying how these factors shape dialogue coherence and emotional realism.

Quality Evaluation and Filtering

Our evaluation process combines human annotation with automated filtering to ensure high-quality conversations. We begin by having three human annotators rate a subset of 984 dialogues, sampling 24 from each domain and 123 per turn length. Annotators evaluate dialogue coherence, emotional consistency, and domain relevance using a Likert scale from 1 to 5, alongside a binary validity judgment.

Table 5.5 **Agreement with human annotations.** Cohen's Kappa (for decision scores), Kendall's Tau, and Spearman correlation (for goodness scores) between human and LLM-based annotations. Best results within each category (closed-source, open-source, ensemble) are underlined, while the overall best are shown in **bold**. The model ultimately used for final filtering is highlighted in light-blue. The table also reports the percentage of dialogues retained after filtering.

Model Type	Model Name	Cohen Kappa (Decision)	Kendall-Tau (Goodness)	Spearman (Goodness)	% Good Dialogues (GT: 59%)
Closed	GPT-3.5-Turbo	0.47	0.37	0.51	76%
	GPT-4o	0.78	**0.64**	**0.79**	54%
	GPT-4o-mini	0.64	0.52	0.66	73%
	Gemini-2.0-Flash	0.61	0.58	0.70	49%
	Gemini-2.5-Flash	0.65	0.62	0.76	46%
	Gemini-2.5-Pro	0.56	0.60	0.77	65%
Open	Phi4	0.39	0.49	0.59	85%
	Gemma3-27B-it	0.68	0.62	0.75	51%
	Llama-3.3-70B-it	0.72	0.60	0.71	65%
	Qwen2.5-72B-it	0.76	0.55	0.70	63%
	Qwen1.5-110B-Chat	0.35	0.54	0.64	86%
Ensemble	Llama-3.3-70B-it + Qwen2.5-72B-it + Phi-4	0.71	0.56	0.75	69%
	Llama-3.3-70B-it + Qwen2.5-72B-it + Qwen1.5-110B-Chat	0.70	0.57	0.75	69%
	Llama-3.3-70B-it + Qwen2.5-72B-it + Gemma3-27B-it	**0.79**	0.62	**0.79**	60%

When assigning negative scores, annotators must specify whether issues stem from incoherence, emotional inconsistency, domain drift, model hallucinations, or other problems. This process achieves substantial inter-annotator agreement with a Fleiss' kappa [301] of 0.80.

These human annotations serve as calibration data for our automated filtering system. We benchmark several state-of-the-art LLMs to identify which best approximates human judgment, testing both open-source models (Phi-4 [297], Gemma-3-27B-Instruct [299], Llama-3.3-70B-Instruct [295], Qwen variants [296]) and closed-source alternatives (GPT family [302, 216, 303], Gemini variants [304–306]). Each model evaluates the human-annotated subset using identical prompts covering the same assessment criteria. Table 5.5 shows the results of the evaluation. Our analysis reveals GPT-4o as the strongest individual model, achieving a Cohen Kappa of 0.78 and Kendall Tau of 0.64 with human judgments. However, we discover that an ensemble combining Gemma-3-27B-Instruct, Llama-3.3-70B-Instruct, and Qwen2.5-72B-Instruct matches or exceeds GPT-4o's performance, reaching a Cohen Kappa of 0.79.

We therefore employ this open-source ensemble to evaluate all 65,600 initial dialogues. Dialogues must achieve scores of 3 or higher across all dimensions and receive positive binary judgment to be retained. We incorporate additional safety filtering using Llama Guard [295] to screen for harmful or toxic content. This rigorous process yields 40,150 high-quality dialogues, approximately 61% of the initial generation, with an average length of 6.1 turns.

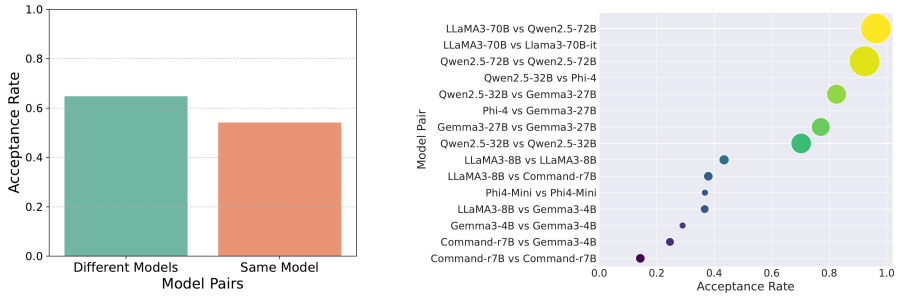
Speech synthesis

To create a truly multimodal resource, we synthesize emotionally expressive speech for all 40,150 dialogues using two complementary approaches. The first leverages XTTS-v2 [307], which enables zero-shot reference voice conditioning.[7] We map our 20 emotional states onto the 8 emotions available in the RAVDESS dataset [95], using reference recordings from 24 different actors. For instance, emotions like "*excited*," "*amused*," and "*enthusiastic*" map to RAVDESS's "*happy*" category, while "*anxious*" maps to "*fearful*." We create reference samples by concatenating standard RAVDESS sentences, using normal intensity recordings to avoid over-exaggerated emotions.

Our second approach employs Orpheus [308], a state-of-the-art TTS model that produces highly natural speech without explicit emotional conditioning.[8] This model relies on linguistic cues within the text itself, such as punctuation and word choice, to generate appropriate prosodic variations. Orpheus provides 8 distinct voices, which we consistently assign to speakers throughout each dialogue.

For both synthesis methods, we maintain speaker consistency by assigning the same voice identity throughout each conversation. Text preprocessing removes non-speech symbols while preserving emotionally significant punctuation. The resulting audio corpus includes all 40,150 conversations in both synthesis variants, totaling over 480 hours of audio across more than 240,000 individual turns per variant. This comprehensive spoken dialogue collection enables new research directions in multimodal conversation modeling and emotion-aware speech synthesis.

[7] https://huggingface.co/datasets/SALT-Research/DeepDialogue-xtts
[8] https://huggingface.co/datasets/SALT-Research/DeepDialogue-orpheus



(a) *AR same vs. different model pairs*

(b) *AR by model pairs*

Fig. 5.6 **DeepDialogue, acceptance rate.** Comparison of acceptance rates (AR) between dialogue pairs produced by identical models and those generated by different models.

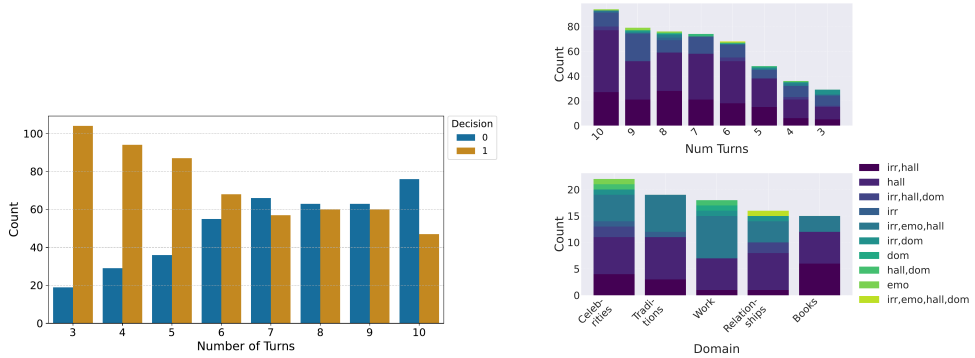
5.4.2 Results and Discussion

Our analysis examines DeepDialogue from two perspectives: an in-depth investigation of dataset properties, and a validation of its emotional content through speech emotion recognition model training.

Empirical Findings from Human Annotations

Our initial evaluation focuses on the human-annotated dialogue subset, with broader dataset analysis available in the original paper.

Model interaction patterns. The impact of model pairing strategies emerges clearly in Figure 5.6. Figure 5.6(a) reveals an evident difference in dialogue acceptance rates between different-model pairs versus same-model conversations. Pairing different models achieves a 0.65 acceptance rate, substantially outperforming the 0.54 rate for identical model pairs, demonstrating a clear "cross-model effect" advantage. Figure 5.6(b) provides deeper insight through a bubble plot of model pair performance. Each bubble represents a unique pairing, with bubble size indicating combined parameter count. The visualization reveals a clear correlation between model size and performance, with larger models consistently clustering toward higher acceptance rates on the right. Pairs including LLaMA3-70B and Qwen2.5-72B show particularly strong performance, while combinations involving smaller models like Command-r7B or Gemma3-4B achieve notably lower acceptance rates.



(a) *Number of valid/invalid dialogues per number of turns.*

(b) *Invalid dialogues, divided by top-5 domains (top), and number of turns (bottom).*

Fig. 5.7 **Distribution of valid and invalid dialogues.** Number of valid and invalid dialogues per number for turns, and reasons for invalidation (b), as evaluated by human annotators.

Impact of conversation length. Figure 5.7(a) reveals how dialogue quality varies with conversation length. Short conversations (≤ 5 turns) show predominantly positive evaluations. At 6 turns, acceptance rates remain higher but with diminishing advantage. Longer exchanges (7-9 turns) see rejection rates approach or exceed acceptances. Ten-turn conversations show a clear majority of rejections. This pattern aligns with recent research [309] highlighting LLMs' deteriorating performance in extended interactions.

Analysis of rejected dialogues. Figure 5.7(b) examines rejected dialogues across domains and lengths. Hallucination (*hall*) and irrelevance (*irr*) emerge as primary failure modes, often coinciding with domain mismatch (*dom*). Emotional inconsistency (*emo*), while less frequent overall, appears more prominently in domains like "*Relationships*" and "*Celebrities*." Longer dialogues (9-10 turns) show higher failure rates, emphasizing the challenge of maintaining coherence in extended conversations. Additional analysis in the original paper confirms smaller models generate most rejections, especially in same-model conversations, while high-capacity models like Qwen2.5-72B and LLaMA3-70B produce substantially fewer invalid dialogues.

Concrete versus abstract domains. To quantify linguistic concreteness, we employ Brysbaert concreteness ratings [310], a lexicon of 40,000 English words scored on a 1-5 scale. Using GPT-4o, we parse accepted dialogues from concrete domains (e.g., "*Cars*") and abstract domains (e.g., "*Philosophy*").

Table 5.6 **Analysis of gender and age biases.** M = male, F = female, Y = young, O = old. PMI denotes the Pointwise Mutual Information scores for gender- and age-specific identity terms. The blocks list the top-10 domains with the largest gender- and age-PMI differences, respectively.

Domain	M/F **ratio**	$\mathbf{PMI}_M$	$\mathbf{PMI}_F$	$\|\mathbf{PMI}_{M-F}\|$	Y/O **ratio**	$\mathbf{PMI}_Y$	$\mathbf{PMI}_O$	$\|\mathbf{PMI}_{Y-O}\|$
Sports	0.00	0.49	-7.44	7.93	1.88	2.47	-1.35	3.82
Coding	0.00	0.49	-6.64	7.13	0.05	-1.23	0.11	1.34
Philosophy	0.01	0.48	-5.64	6.12	0.17	0.28	-0.04	0.32
Makeup	24.00	-4.15	1.74	5.89	0.55	1.59	-0.45	2.04
Holidays	0.01	0.48	-5.03	5.51	0.38	1.23	-0.29	1.52
Technology	0.01	0.48	-4.79	5.27	0.33	1.09	-0.23	1.32
Cars	0.02	0.47	-4.24	4.71	0.10	-0.40	0.05	0.45
Gaming	0.02	0.46	-3.80	4.26	0.11	-0.26	0.03	0.29
Photography	0.02	0.46	-3.59	4.05	0.03	-2.10	0.14	2.24
Shopping	5.53	-2.22	1.55	3.77	0.09	-0.46	0.05	0.51
Finance	0.09	0.37	-1.85	2.22	9.75	2.95	-3.25	6.20
Science	0.00	0.49	0.00	0.49	4.00	2.77	-2.14	4.91
Home	1.76	-0.98	1.15	2.13	0.01	-4.48	0.17	4.65
Sports	0.00	0.49	-7.44	7.93	1.88	2.47	-1.35	3.82
Food	3.03	-1.52	1.38	2.90	0.01	-3.65	0.17	3.82
Cooking	0.96	-0.48	0.77	1.25	0.01	-3.48	0.17	3.65
Education	0.10	0.36	-1.72	2.08	1.67	2.41	-1.23	3.64
News	0.17	0.26	-0.99	1.25	1.00	2.09	-0.82	2.91
Health	0.29	0.12	-0.36	0.48	0.72	1.84	-0.60	2.44
Pets	0.71	-0.28	0.53	0.81	0.71	1.83	-0.60	2.43

Analysis of 1,000 randomly sampled turns per domain type reveals concrete domains average significantly higher concreteness scores (3.01) compared to abstract domains (2.21), as visualized in Figure 5.8.

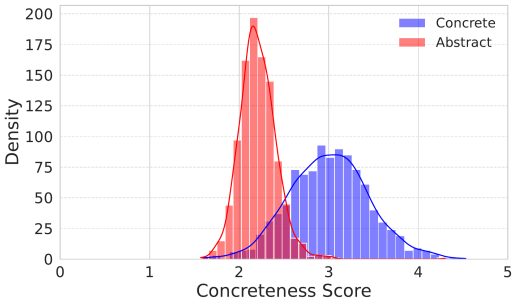


Fig. 5.8 **Comparison of concreteness scores.** Accepted dialogues classified as concrete and abstract via GPT-4o, Brysbaert ratings.

Bias Analysis

We examine gender and age biases using Pointwise Mutual Information (PMI) scores for identity-specific terms across domains.[9] Table 5.6 presents PMI metrics for the 20 most biased domains (10 each

[9]We use the following terms:
`Female`: {*Woman, Girl, Lady, Mother, She, Her, madam, Daughter*}
`Male`: {*Man, Boy, Gentleman, Father, He, Him, Sir, Son*}
`Young`: {*Young, Teen, Child, Kid, Baby*}
`Old`: {*Old, Elder, Senior, Grandpa, Grandma, Aged*}

for gender and age), including male (PMI_M), female (PMI_F), young (PMI_Y), and old (PMI_O) terms, along with absolute disparities and mention ratios. Clear demographic biases emerge across domains. Male-associated terms dominate traditionally masculine fields like *Sports* and *Coding*, showing PMI gaps exceeding 7.0 and sometimes complete absence of female mentions (M/F ratio = 0.00). Conversely, domains like *Makeup* and *Shopping* show strong female bias, with M/F ratios reaching 24.0 and reversed PMI polarity. Age biases manifest similarly, with *Finance* and *Science* skewing toward youth-related language, while domestic topics like *Home* and *Cooking* over-representing older individuals. These patterns suggest LLM-generated dialogues reflect persistent societal stereotypes regarding age and gender roles.

Speech Emotion Recognition

To validate the emotional consistency of our generated dialogues, we explore DeepDialogue's utility for speech emotion recognition tasks. We create a balanced evaluation set from the `XTTS-v2` variant, sampling $1,000$ turns per emotion category. Using this data, we train three self-supervised learning models, i.e., wav2vec 2.0 [1], WavLM [57], and HuBERT [2]. All three models achieve robust performance on DeepDialogue's held-out test set, with both accuracy and macro F1-scores reaching approximately 90% (first block of Table 5.7).

To assess representation transferability, we conduct zero-shot evaluation using the best-performing model on the RAVDESS dataset [95], which shares our emotion label distribution. While we observe an expected performance decrease due to domain shift, our zero-shot HuBERT-DD maintains strong performance with 56.6% accuracy. This approaches the 65.3% accuracy achieved by HuBERT-LP, a linear-probing baseline trained directly on RAVDESS [27]. These results demonstrate that DeepDialogue not only maintains internal emotional consistency but also captures generalizable emotional patterns, confirming its value as a resource for emotion-aware dialogue research.

Table 5.7 **SER.** Speech emotion recognition performance, models trained on DeepDialogue.

Model	**Acc. (%)**	**F1 (%)**
DeepDialogue		
W2V2-DD	88.03	88.10
WavLM-DD	90.20	90.17
HuBERT-DD	93.70	93.64
RAVDESS		
HuBERT-DD	56.64	56.10
HuBERT-LP	65.28	/

5.4.3 Summary and Practical Implications

DeepDialogue is a large-scale multimodal dataset containing $40,150$ high-quality multi-turn dialogues across 41 domains with 20 distinct emotions and coherent emotional progressions. Our comprehensive analysis reveals several key insights about dialogue system behavior: smaller models' performance significantly degrades beyond 6 turns, concrete domains consistently yield superior dialogue quality compared to abstract ones, and cross-model interactions substantially enhance conversational coherence. The dual speech synthesis approach, providing both explicitly emotion-conditioned and implicitly derived emotional expressions, enables novel research at the intersection of text-based dialogue systems and speech-based conversational AI.

The practical applications of DeepDialogue span multiple areas of conversational technology. For virtual assistants, the emotionally coherent dialogues enable development of more engaging and context-aware interactions that maintain appropriate emotional states across multiple turns. In educational applications, the diverse domain coverage and emotional progression modeling can support creation of more natural tutoring systems that adapt their emotional tone to student engagement. The multimodal nature of the dataset makes it particularly valuable for developing healthcare applications where both verbal content and emotional expression must be carefully monitored and responded to.

Despite rigorous quality filtering and demonstrated utility, several important limitations remain. Synthetic dialogues may differ from authentic human interactions in subtle but important ways. Our discrete emotion taxonomy necessarily simplifies the continuous nature of human emotional expression. Current speech synthesis technology cannot fully capture the subtle vocal characteristics that often carry crucial emotional information. Furthermore, our analysis reveals concerning demographic biases across domains, highlighting broader ethical considerations about representation and potential misuse of emotional speech synthesis.

We release DeepDialogue while explicitly acknowledging these constraints, aiming to advance the development of more natural, emotionally intelligent conversational systems. The dataset serves as a foundation for analyzing the dynamics of emotional evolution within conversations and for assessing how different model architectures manage extended interactions. Future work should focus on enhanced emotion modeling to better capture continuous emotional states, expanded cultural representation

to ensure broader applicability, and development of robust safeguards against potential misuse of emotional synthesis capabilities. These advances will be fundamental for creating conversational AI systems that can engage in truly natural, emotionally appropriate, and ethically sound interactions.

5.5 ITALIC: The First Italian SLU Dataset

Effective human-machine interaction requires not only sophisticated emotional modeling (§5.3, §5.4), but also robust language understanding across diverse linguistic contexts. Current spoken language understanding resources heavily favor English, creating significant barriers for speakers of other languages. Although some efforts have addressed multilingual needs [265, 266], existing resources either lack the specificity required for human-machine interaction or miss crucial audio components needed for end-to-end learning.

To address these limitations, we present ITALIC [30], the first large-scale **ITA**lian **L**anguage **I**ntent **C**lassification dataset. Our collection comprises 16,521 audio samples spanning 18 domains and 60 intents, recorded by 70 speakers from various Italian regions. Built upon the Italian portion of the MASSIVE textual dataset [265], ITALIC enriches each recording with comprehensive speaker and channel metadata, including regional origin, demographic information, and recording conditions. This rich annotation enables research beyond intent classification, supporting tasks like speaker recognition, text-to-speech synthesis, and linguistic variety identification.

5.5.1 Data Collection

We collected ITALIC through a web-based crowd-sourcing platform, engaging both native and non-native Italian speakers from diverse regions across Italy. Participants recorded themselves reading instructions randomly selected from the MASSIVE dataset [265], capturing natural language commands typical of virtual assistant interactions. Following initial annotation guidelines, volunteers recorded independently using their own devices without direct supervision, enabling natural variation in recording conditions and environments.

Table 5.8 **Gender and age.** Distribution of gender and age in the ITALIC dataset.

Gender		Age			
Female	Male	[18-25]	[26-40]	[41-55]	≥56
42.96%	57.04%	10.63%	63.86%	10.78%	14.73%

Through an anonymous registration form, participants provided comprehensive demographic and contextual information. We collected age (as a numerical integer), gender identification (male, female, non-binary, or undeclared), region of origin within Italy, and country of origin for non-Italian participants. The form further gathered educational background and documentation of any speech impairments such as lisp or stuttering. For recording conditions, we tracked both device type (laptop microphone, smartphone microphone, or headphones) and environmental noise levels on a scale from completely quiet to very noisy environments. All participants provided explicit consent for their anonymized data and self-declared metadata to be used for research purposes.

To ensure data quality, we implemented a rigorous multi-stage validation process. Each sample required review by at least two independent annotators who evaluated two key criteria: recording intelligibility and coherence with the provided prompt. We conducted multiple validation rounds, beginning with a complete review of the collection, then focusing subsequent rounds specifically on previously identified invalid samples. This iterative process continued until no invalid samples remained in the dataset.

5.5.2 Data Characterization

The final ITALIC dataset contains 16,521 audio recordings totaling 15.46 hours, spanning 60 distinct intents extracted and annotated from the Italian split of MASSIVE. As shown in Table 5.8, the gender distribution shows 42.96% female and 57.04% male participants, with age concentrations primarily in the 26-40 range (63.86%), followed by participants 56 and older (14.73%), 41-55 (10.78%), and 18-25 (10.63%).

Of the 70 distinct volunteers, all but one identified as native Italian speakers, and only four reported speech impairments. The native speakers represent 13 different

Table 5.9 **ITALIC dataset statistics.** Number of utterances, total audio duration (hours), and number of speakers for each split across the three dataset configurations: *massive*, *speaker*, and *noisy*.

Configuration		# Utterances	# Hours	# Speakers
`massive`	train	11514	10.80	69
	validation	2033	1.90	68
	test	2974	2.76	69
`speaker`	train	13123	12.33	56
	validation	1957	1.67	7
	test	1441	1.46	7
`noisy`	train	13742	12.93	69
	validation	1526	1.44	66
	test	1253	1.09	9

Italian regions, capturing the rich linguistic and diatopic variations characteristic of Italian speech [311].

Figure 5.9 visualizes this regional distribution, highlighting the geographical diversity of the dataset and potential for studying regional speech variations.

Recording durations range from 1.14 to 38.34 seconds, with an average length of 3.37 seconds. All audio is encoded in WAV format at a $16kHz$ sampling frequency, ensuring consistent quality across the collection.

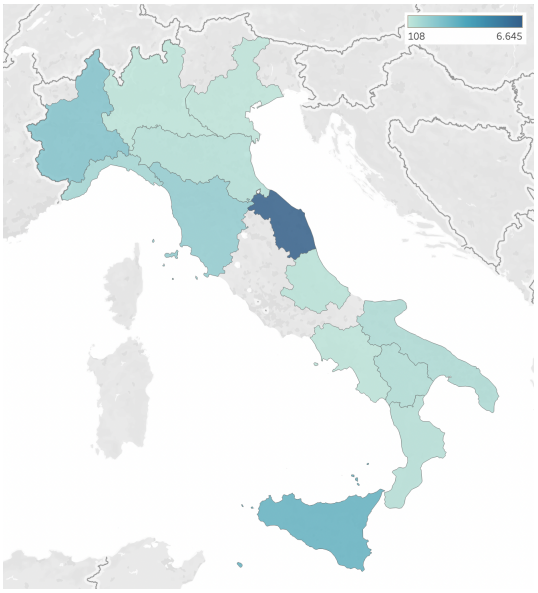


Fig. 5.9 **Utterance distribution across Italian regions.** Darker shades indicate higher counts.

5.5.3 Dataset Splits and Supported Tasks

Dataset splits. To enable robust evaluation across different scenarios, we provide three splitting configurations of increasing difficulty, as detailed in Table 5.9.

The *massive* split follows MASSIVE's official training and test partitions, including all speakers and maintaining a random distribution of noisy utterances between partitions. This configuration enables direct comparison with existing benchmarks on the MASSIVE dataset.

The *speaker* split implements strict speaker stratification, ensuring all recordings from an individual appear exclusively in either training, validation, or test splits. This configuration enables proper assessment of model generalization to completely unseen speakers, a crucial capability for practical applications.

The *noisy* split specifically targets robustness to acoustic conditions by reserving highly noisy data for testing while maintaining noiseless or low-noise data in training and validation. This challenging setup evaluates the ability of the models to maintain performance under degraded acoustic conditions.

As shown in Table 5.9, each configuration maintains substantial training data while providing meaningful validation and test sets. The *massive* split contains $11,514$ training utterances (10.80 hours), the *speaker* split includes $13,123$ training utterances (12.33 hours), and the *noisy* split comprises $13,742$ training utterances (12.93 hours).

Supported tasks. While our primary evaluation focuses on intent classification and automatic speech recognition, the rich annotations of the ITALIC dataset enable investigation of numerous additional research directions. The comprehensive speaker metadata supports tasks in speaker identification, age estimation, and regional accent classification. The high-quality recordings paired with transcriptions enable research in text-to-speech synthesis and pronunciation modeling. Furthermore, the diverse regional representation allows investigation of linguistic variations across Italian dialects. Through this design, ITALIC provides a foundation for advancing Italian spoken language understanding across multiple dimensions of speech processing research.

5.5.4 Experimental Setup

Model selection. We employ transformer-based models that enable end-to-end processing for both intent classification and automatic speech recognition. The IC experiments use either raw audio signals or text transcripts as input, allowing comparison between speech and text-based approaches. Complete implementation

details, including model configurations, hyperparameter settings, and fine-tuning protocols, are available in our project repository.[10] All intent classification models are augmented with a final classification layer attached to their encoder architecture. We utilize implementations and pre-trained weights from the HuggingFace transformers library [149] throughout our experiments.

Intent classification. For speech-based intent classification, we evaluate several variants of XLS-R. We test XLS-R 53 [222] and XLS-R 128 [14] across different parameter scales, with XLS-R 53 at 300M parameters and XLS-R 128 at both 300M and 1B parameters. To assess the impact of language adaptation, we create additional baselines by fine-tuning XLS-R 53 300M and XLS-R 128 1B on ASR tasks using the Italian split of Mozilla Common Voice [60].

For text-based intent classification, we evaluate both multilingual and Italian-specific models. Our multilingual baselines include BERT [58] and BART [312], alongside their Italian-specific counterparts [313, 267]. These models benefit from including Italian in their pre-training languages while offering different approaches to text understanding.

Automatic speech recognition. The ASR evaluation employs three variants of Whisper [13], scaling in complexity: the small architecture with 244M parameters, medium with 769M parameters, and large with 1.5B parameters. This range allows us to analyze the impact of model scale on transcription quality.

Data configuration. We evaluate each model across the three splitting configurations of ITALIC: *massive*, *speaker*, and *noisy*. Models are fine-tuned on the respective training splits and evaluated on the corresponding test sets, enabling assessment under different challenging conditions.

Evaluation metrics. For intent classification, we measure both accuracy and macro F1 scores to account for potential class imbalance. ASR performance is evaluated using both WER and Character Error Rate (CER), providing complementary views of transcription quality.

[10] https://github.com/RiTA-nlp/ITALIC

Table 5.10 **ITALIC, E2E-SLU results.** Accuracy and macro F1 scores for E2E-SLU models and their adapted variants (FT: ✓indicates fine-tuning). Best performance for each data split configuration is highlighted in **bold**.

Split	Model	# params	FT	Accuracy	F1
massive	XLS-R 128	300M		76.16	76.11
	XLS-R 128	1B		77.07	77.08
	XLS-R 53	300M	✓	81.34	81.31
	XLS-R 128	1B	✓	**83.39**	**83.25**
speaker	XLS-R 128	300M		73.42	73.04
	XLS-R 128	1B		79.11	79.08
	XLS-R 53	300M	✓	83.69	83.62
	XLS-R 128	1B	✓	**84.18**	**84.05**
noisy	XLS-R 128	300M		78.29	78.21
	XLS-R 128	1B		76.48	76.06
	XLS-R 53	300M	✓	81.01	80.94
	XLS-R 128	1B	✓	**82.20**	**82.43**

5.5.5 Results and Discussion

Our experimental evaluation focuses on assessing model performance on ITALIC across two key tasks: intent classification and automatic speech recognition. We specifically examine model accuracy, noise robustness, and generalization capabilities to unseen speakers. This investigation also explores how Italian language knowledge and recording condition variations influence model performance.

Speech-based intent classification. Table 5.10 presents the performance of speech SLU models across different ITALIC configurations. Language adaptation proves crucial for performance, with Italian ASR fine-tuning consistently improving results across all configurations. The adapted XLS-R 128 1B model achieves the best performance throughout, showing particularly strong gains of +6.17 and +6.37 F1 points on the *massive* and *noisy* configurations respectively, compared to its non-adapted version.

Model scale generally correlates with improved performance, though the impact is less pronounced than language adaptation. This pattern holds across configurations, with one notable exception: XLS-R 128 300M on the *noisy* split, suggesting

Table 5.11 **ITALIC, NLU results.** Accuracy and macro F1 scores on the Massive split for text NLU models with monolingual (PT: I) and multilingual (PT: M) pre-training. Best performance is highlighted in **bold**.

Model	# params	PT	Accuracy	F1
BART	611M	M	87.16	83.53
BERT	167M	M	86.21	82.93
BART	141M	I	86.65	83.82
BERT	110M	I	**88.43**	**85.57**

that smaller models might sometimes exhibit unexpected robustness to acoustic degradation.

Most models demonstrate stronger performance on the *speaker* configuration compared to *massive*, with only XLS-R 128 300M breaking this trend. This unexpected result suggests these architectures can effectively generalize across different speakers and speaking styles. As anticipated, the *noisy* configuration proves most challenging, with all models except XLS-R 128 300M showing performance degradation, highlighting the significant impact of recording conditions on model reliability.

Text-based intent classification. Table 5.11 presents text NLU model performance on the *massive* configuration. We focus on this split as *speaker* and *noisy* configurations specifically target speech-based evaluation. The Italian pre-trained BERT model achieves superior performance despite its smaller parameter count, outperforming both larger BART architectures and multilingual models. This result reinforces our finding from speech models that language-specific training can outweigh raw model scale. The performance gap between models remains relatively small, suggesting that even compact models can perform well when properly trained on the target language.

Given that all ITALIC samples underwent dual human validation, we can attribute performance differences between speech and text models to the inherent complexity of their respective input modalities. The best text model, monolingual BERT 110M, outperforms the best speech model by 2.32 F1 points despite having 9$\times$ fewer parameters. This gap persists across model pairs, underscoring the additional challenges in processing raw speech compared to text, even with significantly larger models.

Automatic speech recognition. While not primarily designed for ASR evaluation, the diverse speaker pool and recording conditions in ITALIC provide valuable

Table 5.12 **ITALIC, ASRresults.** WER and CER scores for Whisper in either a zero-shot setup (S: ZS) or after fine-tuning (S: FT). Best performance for each data split configuration is highlighted in **bold**.

Split	Model	# params	S	WER	CER
massive	large	1.5B	ZS	11.46	5.01
	small	244M	FT	4.82	1.49
	medium	769M	FT	3.41	0.92
	large	1.5B	FT	**3.06**	**0.82**
speaker	large	1.5B	ZS	8.65	3.93
	small	244M	FT	3.81	0.99
	medium	769M	FT	2.92	0.70
	large	1.5B	FT	**2.74**	**0.61**
noisy	large	1.5B	ZS	15.41	7.67
	small	244M	FT	8.46	2.95
	medium	769M	FT	5.83	1.92
	large	1.5B	FT	**5.29**	**1.70**

insights into Italian ASR performance. Table 5.12 presents results for Whisper models in both zero-shot and fine-tuned configurations across different model scales.

All evaluated models demonstrate strong performance, with metrics approaching the estimated human WER of 4% [314]. Model scale again proves beneficial, with Whisper large consistently outperforming smaller variants across all configurations. The *noisy* split reveals clear performance degradation, particularly pronounced in smaller models, quantifying the relationship between model capacity and noise robustness.

Zero-shot performance lags significantly behind fine-tuned variants, with the gap varying by configuration. The difference is most pronounced on *noisy* and smallest on *speaker*, suggesting that zero-shot generalization struggles most with acoustic degradation. The zero-shot Whisper large achieves 8.65 WER on *speaker* but degrades to 15.41 on *noisy*, performing notably worse than established Italian benchmarks like Mozilla CV (WER: 7.1) [60] and Google Fleurs (WER: 4.0) [273].

5.5.6 Summary and Practical Implications

ITALIC is the first large-scale Italian audio dataset specifically designed for intent classification. Through comprehensive evaluation of state-of-the-art speech and text models, we demonstrated several key findings: language adaptation proves more important than model scale, text models maintain efficiency despite smaller architectures, and acoustic conditions significantly impact performance across all model types. The dataset's diverse speaker pool and recording conditions make it a valuable benchmark for Italian speech technology development.

The practical applications of ITALIC extend across multiple domains of Italian language technology. For virtual assistants and smart home devices, the dataset enables development of more accurate and culturally appropriate command understanding systems. In customer service applications, the regional diversity helps create more robust systems capable of handling various Italian accents and speaking styles. The comprehensive intent coverage supports development of accessible technologies for Italian speakers, while the detailed recording condition annotations aid in creating systems robust to different acoustic environments.

Several important considerations guide future development and application. Regional representation, though broad, does not fully capture the rich Italian landscape of language varieties. The predominance of native speakers limits insight into non-native accent handling, crucial for practical applications. Despite efforts to include diverse speaker profiles, representation of non-binary individuals and those with speech impairments remains limited. The single recording per sentence also constrains analysis of speaker-dependent variations, particularly relevant for SLU applications.

By releasing the dataset, annotation schema, and evaluation code, we provide a foundation for advancing Italian speech understanding research and applications. Future work should focus on expanding speaker diversity and regional coverage to better serve Italy's linguistically diverse population. While achieving a comprehensive representation of the whole linguistic Italian scenario remains challenging, ITALIC establishes an important groundwork for creating more inclusive and effective Italian speech technology.

5.6 Speech Language Varieties in Italy

The analysis of ITALIC reveals an important research gap in modeling the rich linguistic diversity of Italy. Centuries of cultural and historical influences have created a complex assortment of regional language variations, presenting both unique challenges and opportunities for speech technology. Understanding and modeling these regional variations is essential for developing speech systems that serve Italy's linguistically diverse population.

This section explores the automatic identification of Italian language varieties directly from speech signals, without relying on intermediate textual transcriptions [31]. Rather than treating these regional forms as dialects of Standard Italian, we recognize them as distinct language varieties that have developed locally within different geographical areas [315, 316]. This nuanced approach acknowledges the unique linguistic landscape of Italy, where regional languages represent independent developments rather than mere variations of the standard language [317].

5.6.1 Methodology

Our approach combines modern speech representation learning with contrastive objectives to capture subtle distinctions between regional varieties. We leverage pre-trained multilingual models as a foundation, then enhance their ability to discriminate between regional variations through carefully designed contrastive learning strategies. This section details two key components of our methodology: the contrastive learning objectives used to improve representation quality, and the fine-tuning process for adapting pre-trained models to region identification.

Contrastive learning objectives. Contrastive learning improves representation quality by comparing and contrasting positive and negative examples. The model learns class-discriminative features by maximizing similarity between representations of positive examples while minimizing similarity for negatives. We evaluate three supervised contrastive loss functions: supervised contrastive loss (SC), triplet margin loss (TM), and multi-similarity loss (MS).

Supervised Contrastive Loss (SC). The supervised contrastive loss [318] directly optimizes embedding similarities between positive and negative pairs, maximizing agreement between samples from the same region while minimizing agreement

between samples from different regions. Given an encoder network $f(x_i)$ that generates an embedding $z_i = f(x_i)$ for each audio sample $x_i \in X$, the SC loss is defined as:

$$\mathscr{L}_{SC} = -\sum_{i \in I} \frac{1}{|P|_i} \sum_{p \in P_i} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{n \in N_i} \exp(z_i \cdot z_n / \tau)} \tag{5.2}$$

Here, I represents all samples in a batch, P_i denotes positive samples (same region), and N_i represents negative samples (different regions). Following [318], we set the temperature parameter τ to 0.1.

Triplet Margin Loss (TM). The triplet margin loss [319] operates on sample triplets rather than pairs. For an anchor sample x_a, a positive sample x_p, and a negative sample x_n, the TM loss is:

$$\mathscr{L}_{TM} = \max(0, d(z_a, z_p) - d(z_a, z_n) + \mu) \tag{5.3}$$

We use L2 distance for $d(\cdot, \cdot)$ and set the margin parameter μ to 0.05. Triplets are generated using all combinations within each batch where anchor and positives share the same region label.

Multi-Similarity Loss (MS) The multi-similarity loss [157] adaptively selects and weights pairs based on their similarities. It considers the similarity of each sample to itself, other same-class samples, and different-class samples:

$$\mathscr{L}_{MS} = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{\alpha} \log[1 + \sum_{p \in P_i} e^{-\alpha(S_{ip} - \lambda)}] + \frac{1}{\beta} \log[1 + \sum_{n \in N_i} e^{\beta(S_{in} - \lambda)}] \tag{5.4}$$

Where m is batch size, P_i and N_i denote positive and negative samples, S_{ip} and S_{in} are similarities between pairs, and α, β, λ control pair weighting.

We investigate these contrastive objectives through several complementary approaches: using the contrastive loss as the sole objective during additional pre-training, incorporating it as an auxiliary term during fine-tuning, or combining both approaches by first applying contrastive pre-training followed by joint optimization during fine-tuning.

Model Fine-tuning. We adapt pre-trained models to region identification using the VIVALDI dataset (detailed in Section §5.6.2). Most models generate frame-level embeddings from 20*ms* audio segments. We apply average pooling across frames to obtain a single embedding per recording $e = \frac{1}{T} \sum_t e_t$, where e_t represents

frame embeddings and T is the total number of frames. While other methods of obtaining recording-level representations (e.g., attention pooling) could be explored, investigating alternative pooling strategies lies outside the scope of this work but may be considered in future explorations seeking improved performance.

Models are trained end-to-end by minimizing cross-entropy loss between region predictions and reference labels. When using contrastive learning, we optionally include the contrastive objective to simultaneously optimize for classification accuracy and representation quality.

5.6.2 The VIVALDI Dataset

Our analysis leverages VIVALDI (Vivaio Acustico delle Lingue e dei Dialetti d'Italia) [320], a comprehensive collection of spoken utterances capturing Italy's linguistic diversity.[11] The dataset includes recordings from 19 of Italy's 20 administrative regions, excluding only Marche [321–323]. While most regions are represented by samples from three or more cities, Lazio and Campania contribute data from just one city each. Despite this geographic variation, VIVALDI represents the only available large-scale collection of local language recordings spanning such a broad range of Italian locations.

Each city in the collection contributes approximately 343 sentence recordings, with precise geolocation data for each sample. Most cities are represented by a single speaker, with Aidone (Sicily) being the only exception, featuring two speakers. While the absence of speaker demographic information prevents analysis of individual factors, the extensive geographical coverage of the dataset enables robust analysis of regional linguistic variations.

Dataset organization. Table 5.13 presents the dataset statistics across training, validation, and test splits, detailing sample counts, recording duration, city coverage, and average samples per city. As VIVALDI is not provided as a unified collection and lacks predefined splits, we carefully designed the partitioning to ensure robust evaluation. Our splitting strategy enforces two key constraints. First, each city (and consequently each speaker) appears in only one split, preventing models from relying

[11] While ITALIC provides regional speaker information, its recordings are in Standard Italian rather than local varieties, as it was designed for intent classification rather than dialectal analysis. We thus decided to opt for VIVALDI, which also covers a greater number of regions.

Table 5.13 **VIVALDI dataset statistics.** Overview of the VIVALDI dataset splits, including the number of samples, total duration (in minutes), number of represented cities, and the average number of samples per city.

Split	# Samples	# Minutes	# Cities	Avg # per city
Train	81279	2253.88	237	344.40
Val	8242	228.98	24	343.42
Test	8241	227.52	24	343.38

on speaker identification rather than linguistic features. Second, each split must contain samples from at least one city per region to enable proper evaluation across all linguistic varieties. This second constraint led to the exclusion of regions with fewer than three cities, resulting in a final dataset covering 17 of Italy's 20 regions. While this reduced coverage from our initial target of an $80/10/10$ split, it ensures a more challenging and meaningful evaluation where models must generalize to unseen cities within each region, accounting for local linguistic variations.

5.6.3 Experimental Setup

Our evaluation[12] focuses on two key metrics: accuracy and macro F1 score. All reported results represent the mean and standard deviation across three independent runs. We evaluate several state-of-the-art speech models on the VIVALDI dataset, including WavLM [57] as a English-centric baseline, and XLS-R 53 [222] and XLS-R 128 [14] as multilingual models. We also include ECAPA [324], a CNN-based model specifically trained for language identification on VoxLingua107 [325]. Additionally, we evaluate Italian-adapted versions of XLS-R 53 and XLS-R 128 (XLS-R 53-ITA and XLS-R 128-ITA), fine-tuned on CommonVoice's Italian subset [60]. These adaptations aim to better capture Italian linguistic variations.

5.6.4 Results and Discussion

Model performance comparison. Table 5.14 presents the performance of different models using standard fine-tuning for region classification. XLS-R 53-ITA emerges as the best performing model, achieving a macro F1 score approaching

[12] https://github.com/MorenoLaQuatra/SALVI

Table 5.14 **Model selection results.** Mean±std over three independent runs using standard fine-tuning. The ITA-FT column indicates whether models were fine-tuned for Italian ASR (✓) or not (✗). Best-performing models are highlighted in **bold**.

Model	ITA-FT	Accuracy	F1 Macro
WavLM-L	✗	53.35±1.62	43.76±1.14
XLSR-53	✗	56.99±0.61	48.02±1.13
XLSR-128	✗	52.85±2.03	44.95±2.28
XLSR-53-ITA	✓	**60.18±0.55**	**49.84±0.57**
XLSR-128-ITA	✓	55.62±2.24	47.83±2.33

50%. Interestingly, the base XLS-R 53 without Italian fine-tuning outperforms both XLS-R 128-ITA and its pre-trained variant. This unexpected result is likely due to the extensive multilingual training of XLS-R 128, which may limit its ability to capture fine-grained, language-specific nuances. The limited multilingual exposure of WavLM proves insufficient for this specialized task, while ECAPA, despite its language identification focus, performs poorest with macro F1 below 20%.

Impact of contrastive learning. Table 5.15 demonstrates how different contrastive learning strategies affect the performance of XLS-R 53-ITA. We evaluate three approaches: using contrastive loss during pre-training (Ctr-PT), as an additional fine-tuning objective (Ctr-FT), or during standard classification fine-tuning (Clf-FT).

Multi-similarity and triplet margin losses consistently improve performance across all configurations. The multi-similarity objective achieves best overall results, reaching 51.29% macro F1 in multi-task fine-tuning. This strong performance can be attributed to its adaptive pair weighting mechanism, which emphasizes informative contrasts and enhances the ability of the model to capture subtle linguistic distinctions. Conversely, the supervised contrastive loss consistently degrades performance, particularly when used during pre-training.

The most effective strategy employs contrastive loss solely during multi-task fine-tuning. Earlier application through pre-training proves less effective, suggesting the importance of jointly optimizing classification and contrastive objectives specifically for the target task.

Regional correlation analysis. Our t-SNE [159] visualizations in Figure 5.10 reveal how different objectives shape the representation space of the model. The base XLS-R 53-ITA shows no clear clustering structure (Figure 5.10(a)), while contrastive

Table 5.15 **VIVALDI, models performance.** Mean±std over three runs evaluating pre-training (**Ctr-PT**) and fine-tuning strategies (**Ctr-FT** and **Clf-FT**) with different contrastive loss functions on the best-performing model. Best results are highlighted in **bold**.

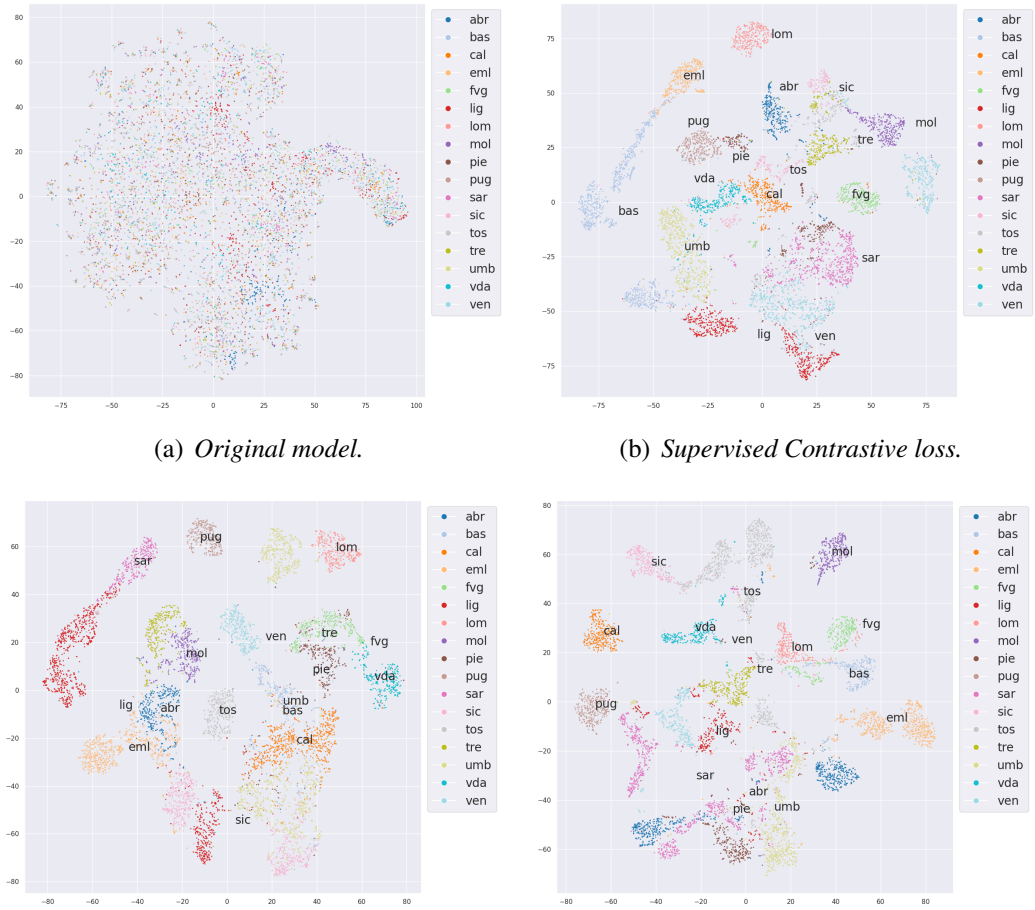
Ctr-PT	Ctr-FT	Clf-FT	Accuracy	F1 Macro
✗	✗	✓	60.18±0.55	49.84±0.57
Supervised Contrastive Loss				
✗	✓	✓	59.02±1.26	49.31±1.33
✓	✗	✓	58.98±0.67	48.82±0.44
✓	✓	✓	57.46±2.10	47.53±1.44
Triplet Margin Loss				
✗	✓	✓	60.23±1.53	50.68±1.71
✓	✗	✓	59.87±0.58	50.56±0.50
✓	✓	✓	58.19±0.64	49.92±1.14
Multi-Similarity Loss				
✗	✓	✓	**60.49±0.88**	**51.29±1.36**
✓	✗	✓	58.92±1.35	51.07±0.61
✓	✓	✓	59.86±0.83	50.98±0.35

pre-training induces more distinct regional groupings. Supervised contrastive loss produces the most overlapping clusters (Figure 5.10(b)), while triplet margin and multi-similarity losses achieve clearer separation (Figures 5.10(c) and 5.10(d)), aligning with their superior quantitative performance.

Figure 5.11 provides deeper insight into our best model's behavior through its confusion matrix (Figure 5.11(a)) and embedding structure (Figure 5.11(b)). The confusion matrix reveals strong correlation between predicted and true regions, though with notable confusion between geographically proximate regions like Basilicata-Calabria and Emilia Romagna-Lombardia.[13] Unexpected confusions, such as Sicilia-Trentino-Alto Adige, may arise from varying degrees of Standard Italian influence in the recordings.

The t-SNE visualization (Figure 5.11(b)) demonstrates the improved discriminative ability of our model compared to both the base variant and contrastive pre-training alone. While most regions form distinct clusters, some overlap persists, reflecting

[13] Region acronyms are provided in our repository.



(a) *Original model.* (b) *Supervised Contrastive loss.*

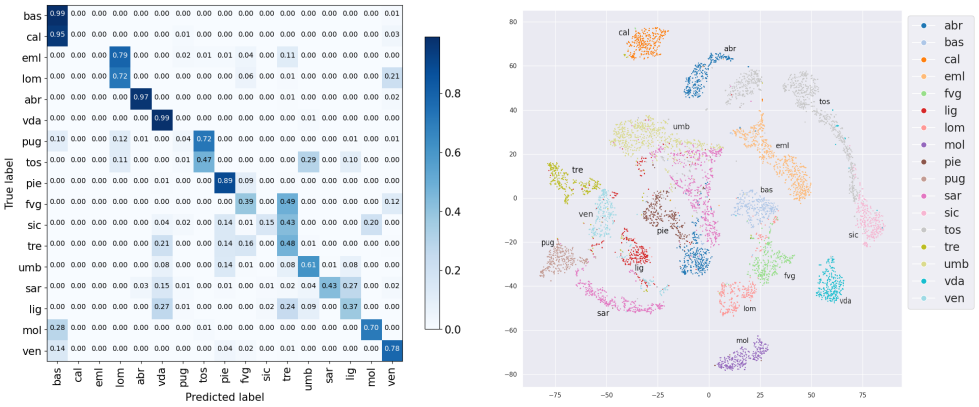
(c) *Triple loss.* (d) *Multi-similarity loss.*

Fig. 5.10 **t-SNE visualization.** Original `XLS-R 53-ITA` model (a) compared to its pre-trained variants using three contrastive learning objectives: supervised contrastive loss (b), triplet-margin loss (c), and multi-similarity loss (d).

the challenge of disentangling subtle variations between certain language varieties. This aligns with previous findings in textual region identification [270], suggesting consistent challenges across modalities in distinguishing closely related linguistic variations.

5.6.5 Summary and Practical Implications

Our analysis demonstrates the feasibility of automatically identifying Italian language varieties directly from speech signals. The combination of multilingual pre-training and contrastive learning proves effective, with XLS-R 53-ITA achieving the strongest



(a) *Confusion matrix.* (b) *t-SNE visualization.*

Fig. 5.11 **Confusion matrix and t-SNE projection of best model.** Confusion matrix (a) and t-SNE visualization (b) of the `XLS-R 53-ITA` model fine-tuned with a multi-task setup using the multi-similarity contrastive objective.

performance when enhanced with multi-similarity contrastive objectives during fine-tuning. While the model successfully captures broad regional distinctions, its confusion patterns reveal the inherent challenges in disambiguating geographically proximate varieties and handling varying degrees of Standard Italian influence.

The practical applications of this work extend beyond mere dialect classification. For educational technology, these models could help develop region-aware language learning tools that respect and preserve local linguistic heritage. In speech synthesis applications, understanding regional variations could enable more authentic and locally appropriate voice generation. The ability to automatically identify language varieties also has important implications for sociolinguistic research, enabling large-scale studies of language variation and change across Italy.

Our findings reveal important considerations for deploying such systems in practice. The model's stronger performance on geographically distant varieties suggests its utility for broad regional categorization, while its limitations with closely related varieties indicate the need for more nuanced approaches in border regions. The observed confusion patterns also provide valuable insights for designing more robust speech processing systems that can handle Italy's complex linguistic landscape.

Looking forward, this work establishes important foundations for preserving and working with Italy's rich linguistic heritage in the digital age. While further research is needed to capture more subtle variations between closely related language varieties,

our approach demonstrates how modern machine learning techniques can help document, analyze, and maintain regional linguistic diversity. This has broader implications for cultural preservation, as similar approaches could be adapted to other linguistically diverse regions facing similar challenges in the digital transformation of their cultural heritage.

5.7 Speech XAI: Making Speech Models Interpretable

Building on our advances in speech understanding and multilingual modeling, we now address a critical challenge that spans all speech processing applications: model interpretability. As speech systems become more sophisticated and widely deployed in real-world scenarios, understanding their internal decision-making processes becomes increasingly important for ensuring reliability and trust. While explainable AI (XAI) techniques have advanced significantly for vision, text, and structured data [326–328], explanations for SLU models remain largely unexplored. Existing approaches focus on spectrogram [274, 275] or phoneme-based [279] explanations, which prove too fine-grained for broader speech tasks where multiple factors like acoustics, linguistics, and prosody interact to convey meaning.

To address these limitations, we propose a novel approach that analyzes speech model decisions at both word and paralinguistic levels [32], providing interpretable insights that align with how humans naturally process and understand speech. This dual-level analysis enables us to capture both the semantic content of utterances and the non-verbal aspects of communication, offering more comprehensive and intuitive explanations of model behavior. By bridging the gap between signal-level analysis and semantic understanding, our approach aims to make speech models more transparent and trustworthy for real-world applications.

5.7.1 Methodology

We present an original approach to explaining speech model predictions by analyzing both linguistic and paralinguistic components of utterances. Our method generates explanations at two complementary levels: word-level audio segments and paralinguistic features. Figure 5.12 demonstrates this dual approach, showing how different utterance components contribute to model predictions.

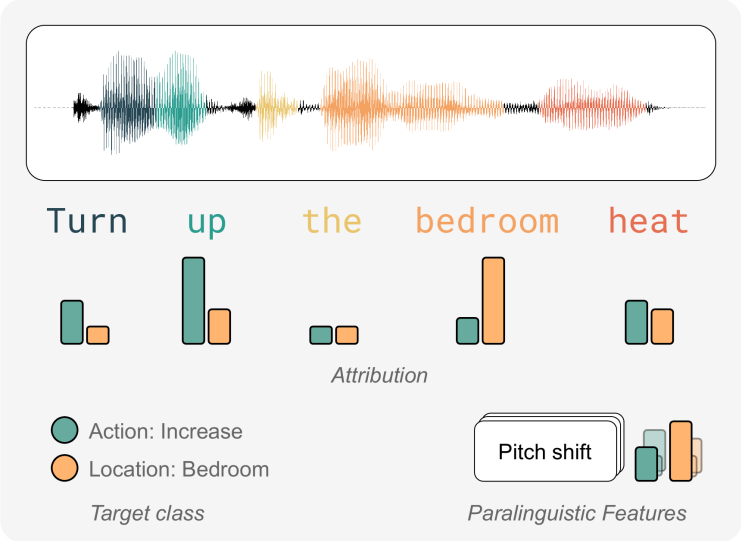


Fig. 5.12 **Word-level and paralinguistic explanation, FSC sample [146].** The audio waveform is aligned with words, color-coded for clarity. Bars indicate word-level attribution strengths for the target classes *Increase* (green) and *Bedroom* (orange).

Word-level audio segment attribution. Our word-level attribution process operates in two phases. First, we perform word-level audio-transcript alignment to extract precise timestamps for each uttered word, defining non-overlapping audio segments in the time domain. When transcripts or timestamps are unavailable, we employ state-of-the-art alignment models [329]. The alignment process segments the audio into non-overlapping word-level regions, excluding silent portions like pauses and signal tails. Since these non-verbal segments carry no explicit semantic content, we exclude them from our attribution analysis.[14] See Figure 5.12 (top) for a visual example of this alignment process.

Second, we compute the contribution of each segment through input perturbation, masking segments by zeroing corresponding samples following prior works [279, 330]. We implement this using two established XAI techniques: Leave-One-Out and LIME [278].

Leave-One-Out attribution. Given an audio signal $x \in \mathbb{R}^n$ with word-level segments $\{x_1,..,x_n\}$, we compute each segment's relevance $r(x_i)$ to predicting class k as:

$$r(x_i) = f(y = k|x) - f(y = k|x \setminus x_i) \tag{5.5}$$

[14] Prosodic features like pause duration are addressed separately through paralinguistic attribution.

where $f(y=k|x)$ is the prediction probability of the model for class k, and $x \setminus x_i$ represents the signal with segment x_i masked. Positive scores indicate segments supporting the prediction, while negative scores suggest opposition.

LIME-based attribution. LIME approximates complex models locally using simpler, interpretable surrogates. We represent inputs using binary vectors indicating masked/unmasked word segments, enabling neighborhood sampling as shown in Figure 5.13.[15] Unlike Leave-One-Out, LIME can mask multiple segments simultaneously, capturing potential interaction effects between words.

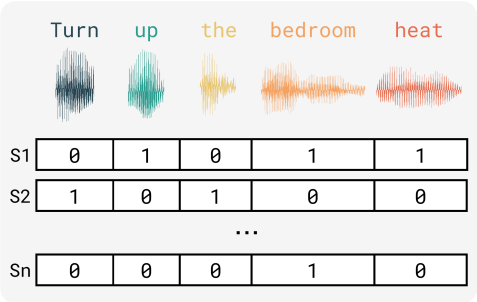


Fig. 5.13 **LIME Example.** Word-level time alignment highlights the audio segments to be masked (top). The LIME sampling process then selects specific segments for masking (bottom), where 1 indicates a masked segment and $S_1, \ldots, S_n$ represent the sampled neighborhood variants.

Paralinguistic attribution. Beyond word-level analysis, we examine the impact of paralinguistic features through targeted signal perturbations. For a paralinguistic measure p (e.g., pitch or signal-to-noise ratio), we compute its relevance by averaging the effect of multiple perturbations:

$$r_p(\widetilde{x}) = f(y=k|x) - f(y=k|\widetilde{x}) \tag{5.6}$$

$$r(x,p) = \frac{1}{|\widetilde{X}_p|} \sum_{\widetilde{x} \in \widetilde{X}_p} r_p(\widetilde{x}) \tag{5.7}$$

where $\widetilde{X}_p$ represents the set of transformed signals. The resulting scores, bounded between -1 and 1, indicate the sensitivity of the model to each paralinguistic feature. For instance, high sensitivity to background noise might indicate vulnerability to acoustic perturbations, while strong pitch dependence could suggest gender or age-related biases in the decision-making process of the model.

[15] While we focus on word-level segments, other approaches like fixed-width windows or n-grams could be explored in future work to capture additional linguistic patterns.

Table 5.16 **FSC sample, word-level explanation of audio segments.** Darker colors and higher values indicate greater relevance of the segment to the model's prediction.

	Turn	**up**	**the**	**bedroom**	**heat.**
act=increase	0.250	0.545	0.260	0.139	0.021
obj=heat	0	0	0	0.014	0.550
loc=bedroom	0.002	0.006	0.087	0.997	0.323

Table 5.17 **Paralinguistic Feature Attribution.** Example of $r(\mathbf{x}, p)$ for p: time stretching, pitch shifting, and noise injection on a sample from the FSC dataset (see Table 5.16). Darker colors and higher values indicate greater model sensitivity to perturbing the feature.

	speed		**pitch**		**reverb**	**noise**
	up	**down**	**down**	**up**		
act=increase	0.19	0.04	0.04	0.13	0.56	0.44
obj=heat	0	0	0	0.04	0	0.29
loc=bedroom	0.03	0.01	0.13	0.33	0.36	0.60

5.7.2 Experimental setup

Datasets and tasks. We evaluate our method[16] across three datasets spanning two tasks. For intent classification we employ FSC [146] in English and ITALIC [30] in Italian (*Speaker* split). For speech emotion recognition we use IEMOCAP [94]. For the latter, we specifically focus on Session '1' with 942 utterances labeled for five emotions: happiness, anger, sadness, frustration, and neutral.

Models and implementation. For English datasets (FSC and IEMOCAP), we employ wav2vec 2.0 base [1] using pre-existing fine-tuned checkpoints [149]. For ITALIC, we use multilingual XLS-R 128 [14] with its corresponding fine-tuned versions [30]. We use WhisperX [329] for word-level alignment and transcription, achieving WERs of 1.72 on FSC, 15.77 on IEMOCAP, and 7.49 on ITALIC. For paralinguistic features, we implement pitch shifting, time stretching, background white noise injection, and reverberation.

5.7.3 Results and Discussion

Our analysis examines both instance-level and global explanations [331], providing complementary views of model behavior. Instance-level analysis reveals how models make specific predictions, while global patterns expose broader behavioral trends.

[16] https://github.com/elianap/SpeechXAI

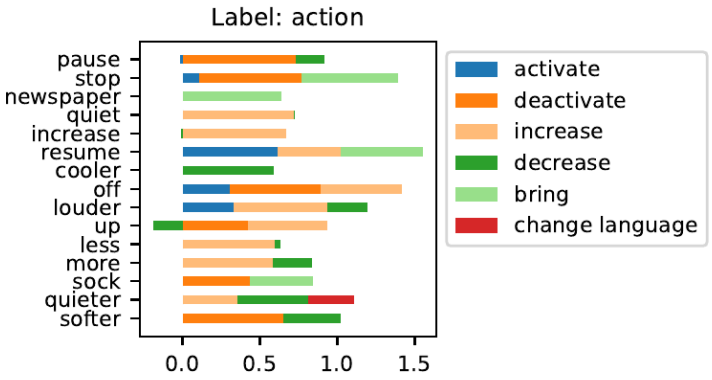


Fig. 5.15 **Top 15 most influential words** for each predicted class in the FSC dataset (Slot: *Action*).

Instance-level analysis. Table 5.16 demonstrates our approach on an FSC example: "*Turn up the bedroom heat.*" The word-level attributions clearly show "*up*" driving the "*increase*" action prediction, while "*heat*" and "*bedroom*" strongly influence their respective slot predictions.

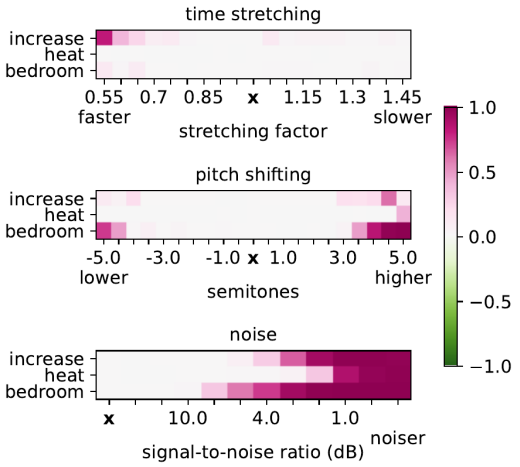


Fig. 5.14 **Paralinguistic Effect Breakdown.** Visualization of $r_p(\widetilde{\mathbf{x}})$ for p: time stretching, pitch shifting, and noise injection. $\mathbf{x}$ denotes the original signal. Darker red indicates a stronger drop in probability.

The paralinguistic analysis in Table 5.17 reveals varying sensitivities. Speed modifications primarily affect the "*increase*" action prediction, pitch changes impact location and action predictions, and noise most strongly influences location classification. Figure 5.14 provides a detailed breakdown of these effects through attribution heatmaps.

Global analysis. To understand broader model behaviors, we aggregate attribution scores across the entire dataset. Figure 5.15 presents the top 15 most influential words for the slot "*action*" in FSC, revealing clear patterns in word-prediction associations. Some words like "*newspaper*" and "*cooler*" strongly correlate with single actions ("*bring*" and "*decrease*", respectively), indicating clear semantic relationships. However, we also identify potential issues, such as "*pause*" being erroneously linked to "*decrease*" while correctly associated with "*deactivate*".

Table 5.18 **Average $r(\mathbf{x}, p)$ for p:** time stretching (speed variation), pitch shifting, and noise injection on the FSC dataset. Darker colors (higher scores) indicate greater model sensitivity to each perturbation.

	speed		pitch		reverb	noise
	up	down	down	up		
action	0.13	0.09	0.12	0.07	0.27	0.37
object	0.07	0.05	0.07	0.04	0.17	0.43
location	0.06	0.04	0.06	0.04	0.11	0.21

Table 5.19 **Comprehensiveness (↑, top) and Sufficiency (↓, bottom) scores.** Word attribution explanations using leave-one-out (WA-L1O), word-level LIME (WA-LIME), and random attribution on the FSC, ITALIC, and IEMOCAP datasets, reported per label. Best results are shown in **bold**.

	FSC			ITALIC	IEMOCAP
Method	**action**	**object**	**location**	**intent**	**emotion**
WA-L1O	0.623	0.627	0.467	0.693	**0.507**
WA-LIME	**0.638**	**0.663**	**0.481**	**0.723**	0.484
random	0.299±0.002	0.251±0.005	0.192±0.005	0.325±0.005	0.274±0.005
WA-L1O	**0.161**	0.086	0.063	0.158	0.310
WA-LIME	0.165	**0.077**	**0.054**	**0.139**	**0.264**
random	0.483±0.004	0.447±0.007	0.338±0.002	0.558±0.002	0.454±0.004

Table 5.18 summarizes the global impact of paralinguistic features. Time compression ("*stretch down*") shows particularly strong effects on action classification. Pitch modifications, especially lowering, significantly influence action predictions. Background noise consistently impacts predictions across all slots, demonstrating the general sensitivity of the model to acoustic degradation.

Quantitative evaluation. Following established XAI evaluation frameworks [332], we assess our explanations through both faithfulness and plausibility metrics. Faithfulness measures how accurately explanations reflect model behavior, while plausibility evaluates alignment with human expectations.

Faithfulness analysis. We adapt comprehensiveness and sufficiency metrics [151] for audio segment explanations. Comprehensiveness measures whether highlighted segments truly drive model predictions, assessed by progressively masking relevant portions. Sufficiency evaluates whether highlighted segments alone maintain model predictions. Table 5.19 compares our Leave-One-Out (WA-L1O) and LIME-based

Table 5.20 **User study on visualization effectiveness.** Mean and standard deviation of participant scores across the four evaluation questions.

	FSC			ITALIC		
Category	Bar	Words	Table	Bar	Words	Table
Identify words (↑)	2.54±0.74	**3.54±0.51**	3.40±0.85	3.54±0.61	3.51±0.51	**3.66±0.54**
Compare word (↑)	3.37±0.77	2.60±0.81	**3.63±0.60**	**3.66±0.68**	2.74±0.82	3.63±0.65
Inspect multiple (↑)	2.57±1.01	**3.34±0.73**	3.29±0.83	**3.34±0.76**	3.29±0.75	3.29±0.62
Overall pref. rank (↓)	2.51±0.70	1.94±0.73	**1.54±0.74**	1.89±0.83	2.20±0.83	**1.80±0.68**

Table 5.21 **User Study, Visualization Effectiveness I.** p-values from the pairwise Wilcoxon test. Statistically significant pairwise differences (p-value <0.05) are shown in **bold**.

	FSC			ITALIC		
Category	Bar, Words	Bar, Table	Word, Table	Bar, Words	Bar, Table	Word, Table
Identify words	**<0.0001**	**0.0001**	0.2291	0.4284	0.1425	0.1126
Compare words	**0.0008**	0.0752	**<0.0001**	**<0.0001**	0.4278	**0.0001**
Inspect multiple	**0.0016**	**0.0010**	0.3867	0.4364	0.3583	0.4762
Overall pref. rank	**0.0083**	**0.0001**	0.0518	0.0802	0.3187	**0.0297**

(WA-LIME) approaches against a random baseline across all datasets. Both methods significantly outperform random attribution, with WA-LIME generally showing stronger faithfulness scores. This suggests our explanations successfully identify genuinely influential audio segments.

Plausibility user study. We conduct a user study with 35 machine learning practitioners to assess explanation plausibility, focusing on intent classification in both English and Italian. The study evaluates both explanation quality and visualization preferences. To validate our approach, participants first compared our explanations against random attribution baselines. We then collect plausibility ratings using a 4-point Likert scale and compared three visualization strategies: our color-coded score approach, traditional word saliency maps, and bar charts.

Table 5.22 **User Study, Visualization Effectiveness II.** p-values from the Friedman test. Statistically significant differences (p-value < 0.05) are shown in **bold**.

Category	**FSC**	**ITALIC**
Identify words	**<0.0001**	0.3577
Compare words	**<0.0001**	**<0.0001**
Inspect multiple	**0.0092**	0.7030
Overall pref. rank	**0.0002**	0.2564

The results (Tables 5.20, 5.21, 5.22) strongly support our method's plausibility. All participants preferred our explanations over random baselines for both FSC

and ITALIC datasets. Our explanations received high plausibility scores: $3.13/4$ ($\sigma = 0.787$) for FSC and $3.37/4$ ($\sigma = 0.75$) for ITALIC. Statistical analysis (p <0.05, Friedman test) showed significant differences between visualization strategies.

Visualization preferences varied by task complexity. For multi-label tasks like FSC, participants preferred our color-coded tables and saliency maps. In single-label tasks like ITALIC, bar charts emerged as equally effective. Score comparison was best facilitated by our color-coded tables and bar charts across all scenarios.

These findings suggest that visualization strategy should adapt to task complexity, with color-coded approaches particularly valuable for multi-label scenarios. The consistently high plausibility scores across languages and datasets demonstrate our method's robustness and utility for model interpretation.

5.7.4 Summary and Practical Implications

Our explainability framework provides the first comprehensive approach to understanding speech model decisions through both linguistic and paralinguistic lenses. By combining word-level audio segment attribution with paralinguistic feature analysis, we enable detailed insights into how models process and combine different aspects of speech input. The strong performance across multiple languages, tasks, and model architectures demonstrates the framework's versatility, while user studies confirm the interpretability of our explanations.

Our findings reveal important patterns in speech model behavior: the varying impact of acoustic modifications across different prediction tasks, the interaction between word-level and paralinguistic features, and the models' sensitivity to specific input perturbations. These insights prove particularly valuable for conversational AI, where understanding how models interpret emotional cues, handle non-verbal vocalizations, and process different language varieties is crucial for creating natural interactions. The ability of our framework to explain decisions across both verbal and non-verbal components aligns directly with our broader goal of developing more human-like conversational systems.

The practical implications extend throughout the conversational AI pipeline. For emotional dialogue systems like those built on DeepDialogue, our framework can help validate whether models correctly interpret emotional cues across turns. In multilingual applications, such as those using ITALIC, it can reveal how models handle

regional variations and accents. For systems processing non-verbal vocalizations with voc2vec, it provides key insights into how different acoustic features contribute to emotional understanding. These insights not only aid in model debugging and improvement but also contribute to developing more transparent and trustworthy conversational systems.

By releasing our framework as open-source software, we provide researchers and practitioners with tools to better understand and validate their speech models' decision-making processes. This transparency is essential for deploying conversational AI in real-world applications where understanding and trust are paramount, from healthcare communications to educational interactions.

5.8 Conclusions

This chapter has presented four complementary contributions advancing speech technology toward more natural, emotionally-aware, and interpretable systems. Each contribution addresses distinct but interconnected challenges in modern speech processing.

Our first contribution, voc2vec (§5.3), establishes a foundation model specifically designed for non-verbal audio processing. By leveraging self-supervised learning on carefully curated non-verbal datasets, voc2vec demonstrates significant improvements in tasks ranging from emotion recognition to baby cry detection. The success of the model in capturing subtle emotional expressions through non-verbal cues provides an important building block for more empathetic speech systems.

DeepDialogue (§5.4) advances the field further by providing a comprehensive multimodal dataset for emotionally coherent conversations. With $40,150$ high-quality dialogues spanning 41 domains and 20 distinct emotions, it enables research into how emotions evolve naturally through multi-turn interactions. Our analysis revealed critical insights about model behavior, including the degradation of smaller models beyond six turns, the benefits of cross-model interactions for dialogue coherence, and inner model biases.

ITALIC (§5.5) addresses the fundamental need for linguistically diverse speech understanding by introducing the first large-scale Italian SLU dataset. Through careful experimental design and comprehensive evaluation, we demonstrated how language-

specific adaptation outperforms pure model scaling, while providing insights into the challenges of handling regional variations and acoustic conditions. Building on this foundation, our detailed analysis of Italian language varieties (§5.6) demonstrated the feasibility of automatically identifying regional speech patterns, while revealing important challenges in distinguishing between geographically proximate varieties.

Finally, our Speech XAI framework (§5.7) completed this progression by making speech model decisions interpretable through both linguistic and paralinguistic lenses. By combining word-level attribution with paralinguistic feature analysis, we enabled detailed understanding of how models process and integrate different aspects of speech input. User studies confirmed the framework's effectiveness in providing meaningful explanations across languages and tasks.

Collectively, these contributions lay the groundwork for developing speech systems that are more natural, emotionally aware, and trustworthy. From capturing non-verbal cues to modeling emotional dialogue progression, from expanding linguistic coverage to preserving regional language varieties, and from making model decisions interpretable, each advance brings us closer to speech technology that can truly understand and engage with the full spectrum of human communication.

5.8.1 Future Research Directions

While our contributions advance natural conversational AI across multiple dimensions, several key challenges remain for future research.

Multimodal integration. Future work should explore how non-verbal vocalizations interact with other communication channels like facial expressions and gestures. This integration is important for developing systems that can process the full spectrum of human communication, including mixed emotions and rapid transitions in conversational dynamics.

Cross-cultural adaptation. Research is needed on efficiently adapting models across languages and cultures while preserving local variations. This includes developing techniques for few-shot learning of new dialects, handling code-switching naturally, and maintaining model robustness across diverse speaking styles and regional varieties. Our work on Italian language varieties highlights both the promise and challenges in this direction, particularly for languages with rich regional diversity.

Advanced interpretability. While our explainability framework provides initial tools, future work should investigate more sophisticated methods for analyzing the interaction between verbal content, prosody, and paralinguistic features across longer temporal contexts. This could help identify underlying factors driving model decisions and enable more effective bias detection and mitigation.

Ethical deployment. As these technologies mature, research must address privacy concerns in emotional speech processing, develop robust consent mechanisms, and establish guidelines for responsible deployment. This includes preventing misuse while ensuring accessibility for legitimate applications.

Progress in these directions will require interdisciplinary collaboration and careful consideration of both technical and societal implications. Success could enable the next generation of conversational AI systems that are not only more capable, but also more trustworthy and equitable. Building on this vision, our contributions have already established fundamental advances in making conversational AI more natural and inclusive: from voc2vec's pioneering work in non-verbal understanding, through DeepDialogue's comprehensive emotional modeling, to ITALIC's expansion of linguistic coverage and our novel approach to speech model interpretability. This work not only pushes technical boundaries but also provides foundational steps toward AI systems that converse naturally, express emotional intelligence, respect cultural context, and operate with transparency and trust.

Chapter 6

Medical Applications of Speech Technology

6.1 Introduction

The previous chapters explored fundamental advances in speech technology across multiple dimensions: from improving model robustness and fairness through bias detection and mitigation techniques (Chapter 3), to developing comprehensive evaluation frameworks that reveal critical failure modes (Chapter 4), to advancing natural conversation modeling through emotional understanding and linguistic diversity preservation (Chapter 5). This chapter examines how these advances can be applied to critical medical challenges, where speech analysis serves both as a diagnostic tool and as an assistive technology. Recent research has demonstrated that voice and speech patterns can serve as important biomarkers for various medical conditions, from voice pathologies to neurological disorders [5, 333–335]. However, developing reliable speech-based medical applications presents unique challenges that push the boundaries of current speech technology.

The impact of voice and speech disorders on quality of life is substantial and widespread [336–342]. Studies indicate that approximately 30% of the general population experiences voice disorders during their lifetime, with significant implications for personal, professional, and social functioning. These conditions range from functional issues like muscle tension dysphonia to organic pathologies such as vocal cord nodules [343–345]. Early detection is crucial, as untreated disorders

often progress to chronic conditions and can cause both physical and psychological distress. Current clinical diagnosis typically relies on specialized equipment and expert clinicians, making it costly, potentially invasive, and limited by specialist availability.

Similarly, motor speech disorders like dysarthria affect millions of individuals worldwide, creating significant barriers to communication and technology access. These conditions, often associated with neurological conditions such as Parkinson's disease, cerebral palsy, or stroke, can make speech difficult to understand even for human listeners [346]. As voice-controlled interfaces become increasingly prevalent in daily life, individuals with speech disorders risk being excluded from these technological advances [347]. This creates an urgent need for more inclusive and adaptive speech recognition systems that can accurately process diverse speech patterns.

The development of speech-based medical applications faces three fundamental challenges that distinguish it from general speech technology development.

First, pathological speech often deviates significantly from the typical patterns that foundation models are trained on. These deviations can manifest in multiple ways: irregular articulation, atypical prosody, inconsistent speaking rates, and various forms of disfluency [348]. Standard speech models, even those achieving impressive performance on typical speech, often fail when confronted with these variations. This challenge is worsened by the fact that pathological speech patterns can vary widely even within the same diagnostic category, making it difficult to develop robust, generalizable solutions.

Second, medical applications demand extremely high accuracy and interpretability. Unlike general speech applications where occasional errors might be merely inconvenient, mistakes in medical contexts could lead to misdiagnosis or inappropriate treatment decisions. This requires not only highly accurate models but also systems that can provide clear explanations for their decisions and quantify their uncertainty. Furthermore, many medical applications require analyzing subtle acoustic features that might be irrelevant for standard speech recognition but fundamental for diagnosis.

Third, medical data is often scarce and sensitive. Privacy regulations and ethical considerations make it difficult to collect large datasets of pathological speech. This scarcity creates challenges for modern deep learning approaches, which typically require substantial amounts of training data. Additionally, the sensitive nature of

medical data requires careful attention to privacy preservation and secure processing methods.

To address these challenges, this chapter presents three complementary contributions that advance the state of medical speech technology.

First, we introduce a transformer-based approach for voice pathology detection that combines acoustic analysis of both sustained vowels and sentence readings [33]. This work addresses the fundamental question of how different types of vocal tasks can provide complementary information for diagnosis. Through careful architectural design and specialized pre-training strategies, our method achieves significant improvements in pathology detection accuracy. We demonstrate that by leveraging multiple sources of vocal information we can create more robust and reliable diagnostic tools.

Second, we develop a novel framework for analyzing clinical speech that leverages multi-source fusion through a multimodal architecture [34]. This approach extends [33] and explicitly models the relationship between different vocal tasks, allowing the system to combine evidence in ways that mirror clinical practice. Our results show substantial improvements in detection accuracy across multiple pathological conditions, with particularly strong performance on challenging cases where single-source analysis might fail.

Third, we present an innovative solution for dysarthric speech recognition that combines end-to-end ASR with generative error correction [35]. This two-stage approach separates the challenges of acoustic modeling from linguistic reconstruction, allowing each component to specialize in its specific task. The system first generates multiple hypotheses about what was said, then uses a large language model to analyze these collectively and produce accurate transcriptions. This approach achieves substantial improvements in transcription accuracy while maintaining computational efficiency, making it practical for real-world deployment.

Beyond their immediate practical applications, these contributions advance our understanding of how modern machine learning techniques can be adapted for specialized medical applications. They demonstrate the importance of carefully considering the unique characteristics of medical data when designing speech technology solutions. Our work shows that by combining insights from clinical practice with state-of-the-art deep learning approaches, we can create systems that are both more accurate and more reliable than traditional methods.

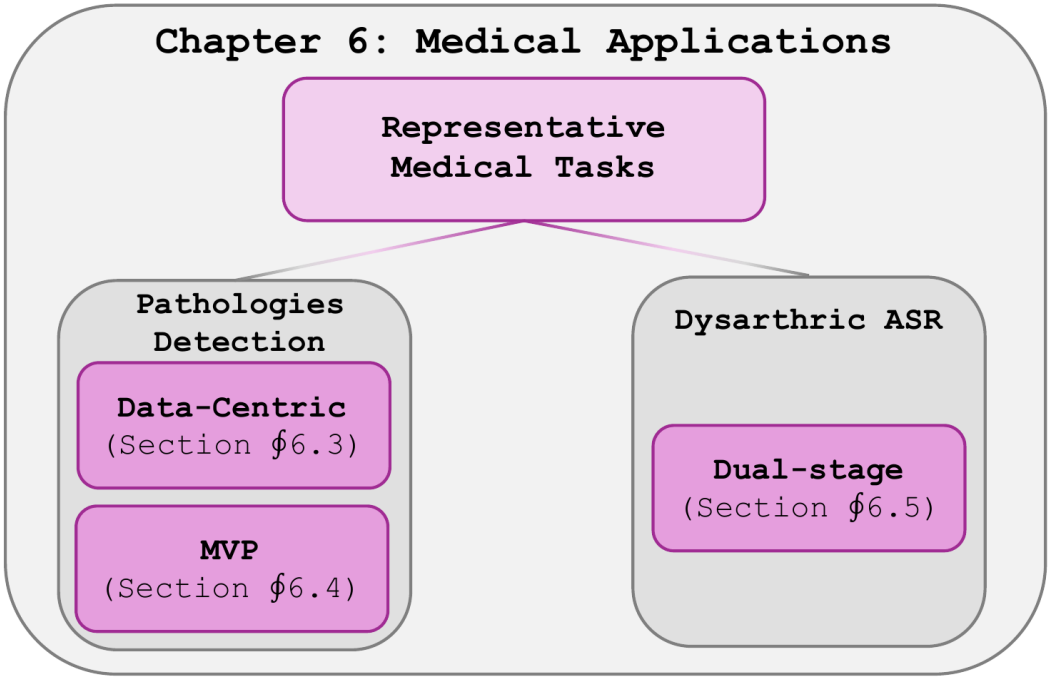


Fig. 6.1 **Chapter 6 Overview.** Graphical taxonomy of Chapter 6 topics.

The findings presented in this chapter have significant implications for the future of medical speech technology. They suggest paths forward for developing more inclusive speech interfaces that can serve individuals with diverse speech patterns. They also provide insights into how speech analysis might be used more broadly as a non-invasive diagnostic tool across various medical conditions. Most importantly, they demonstrate that with careful attention to medical requirements and constraints, modern speech technology can make meaningful contributions to healthcare delivery and accessibility.

The remainder of this chapter is organized as follows. Section §6.2 reviews relevant literature in voice pathology detection and dysarthric speech recognition. Section §6.3 presents our data-centric approach to voice pathology detection using transformer models. Section §6.4 introduces MVP, our framework for multi-source voice pathology detection. Section §6.5 describes our two-stage approach to dysarthric speech recognition combining ASR with generative error correction. Finally, Section §6.6 synthesizes our findings and discusses their implications for medical speech technology. A graphical taxonomy of the contributions presented in this chapter is given in Figure 6.1.

6.2 Related Work

The application of speech technology to medical domains has evolved significantly with advances in machine learning. We review key developments in two representative areas: voice pathology detection (§6.2.1) and dysarthric speech recognition (§6.2.2), focusing on how methodological approaches have adapted to address the unique challenges of clinical applications.

6.2.1 Voice Disorders

Voice disorder analysis has seen significant evolution in methodological approaches over recent years. Early studies focused on feature-based methods, employing multilayer perceptrons trained on extracted features like acoustic parameters and Mel-frequency cepstral coefficients (MFCCs) [349, 350]. These approaches, while providing a foundation for automated analysis, were limited by their reliance on hand-crafted features.

The field advanced with the application of deep learning architectures to spectral representations. Convolutional Neural Networks (CNNs) demonstrated strong performance when applied to 2D representations like Mel spectrograms and MFCC cepstrograms [351, 352]. Hybrid architectures combining CNNs with Recurrent Neural Networks (RNNs) further improved performance by better capturing temporal dependencies in longer voice signals [353].

Recent research has shifted toward end-to-end approaches that process raw audio directly. Models including 1D-CNNs [354] and transformers [6] have shown promising results by learning directly from waveforms. Transformer architectures, in particular, have demonstrated strong performance not only in voice pathology detection but also in related tasks like dysarthric speech analysis [355–357].

While these advances have improved detection capabilities, current methods typically analyze either sustained vowels or continuous speech in isolation [6, 356]. This limitation fails to capture the complementary information available in different vocal tasks. Sustained vowels provide stable conditions for analyzing voice quality but miss the dynamics of natural speech. Conversely, continuous speech captures everyday voice use patterns but introduces complexity through linguistic content and prosody.

The importance of multi-source analysis is highlighted by clinical evidence showing that voice pathologies manifest differently across speaking tasks. Vocal nodules may show more prominent effects during sustained phonation, while muscle tension disorders might be more apparent in continuous speech [358, 359]. This variation in manifestation suggests the potential value of combining multiple sources of vocal information.

6.2.2 Dysarthric Speech Recognition

In parallel, significant advances have been made in speech recognition technology, particularly for challenging conditions like dysarthric speech. Modern ASR systems have achieved remarkable improvements through self-supervised learning and large-scale training [1, 2, 13]. However, these systems still struggle with dysarthric speech, often showing error rates exceeding 30% [348].

Recent approaches to dysarthric speech recognition have focused primarily on acoustic model adaptation [360, 361]. While these methods have shown promise, they may not fully address the complex challenges of pathological speech recognition. The emergence of large language models has opened new possibilities for error correction through text-to-text mapping [362, 363], though this approach remains relatively unexplored in the medical domain.

The development of these technologies is further supported by new resources like the Speech Accessibility Project dataset [348]. Unlike previous collections, this dataset provides both greater scale and more detailed annotations, enabling more comprehensive analysis of pathological speech patterns. These resources, combined with advances in model architecture and training strategies, are enabling new approaches to medical speech technology that better serve individuals with speech impairments.

6.3 Transformers for Voice Pathology Detection: A Data-Centric Approach

Voice disorders significantly impact patient quality of life, yet their automated detection remains challenging due to data limitations. While transformer-based models have shown remarkable success in general speech processing, their application to

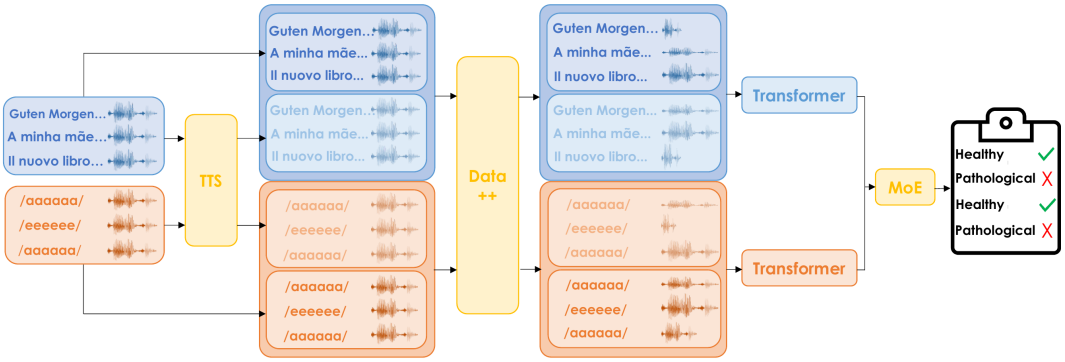


Fig. 6.2 **Schematic overview of the proposed pipeline.** Training data are first synthesized (`TTS`) and augmented (`data++`) for both sentences and sustained vowels. Two specialized `transformer` models are then trained, and their predictions are combined using a shallow Mixture of Experts (`MoE`).

medical voice analysis faces two key constraints. First, these models require substantial amounts of training data, yet medical datasets are typically small and imbalanced. Second, different vocal tasks can provide distinct diagnostic information, requiring careful consideration of how to leverage multiple recording types effectively.

This section presents a novel approach that addresses these challenges through synthetic data generation and a shallow Mixture of Experts (MoE) framework [33]. By combining these two techniques, we demonstrate how transformer models can be effectively adapted for medical voice analysis despite data constraints, showing significant improvements in pathology detection and classification accuracy with respect to current literature.

6.3.1 Methodology

Our approach focuses on maximizing the utility of transformer models for non-invasive voice disorder analysis. While these models typically require large quantities of training data, medical datasets often provide limited samples with uneven representation across diagnostic categories. Public datasets frequently show significant imbalances between healthy and pathological cases, complicating reliable model training.

To address these data limitations, we introduce a synthetic data generation approach conditioned on real patient voices. This process creates additional training samples for both healthy and pathological conditions, helping balance the dataset distribution.

We complement this synthetic data with a comprehensive augmentation pipeline to further enhance the robustness of our models.

Beyond data enhancement, our method innovates by considering multiple types of vocal recordings simultaneously. Different vocal tasks can provide distinct insights into voice disorders. The detection of specific pathologies may manifest differently across various recording types, such as sustained vowels versus read sentences. Unlike previous work that focuses on single recording types [6, 349–352, 354], we employ a shallow Mixture of Experts (MoE) ensemble to combine predictions from models specialized for different vocal tasks. An overview of the entire pipeline is depicted in Figure 6.2.

Synthetic data creation. To generate both healthy and pathological voices, we employ Text-to-Speech (TTS) technology with class-conditional generation. The TTS generation process incorporates learned embeddings derived from the vocal characteristics of real data within each class, ensuring the synthetic samples reflect authentic voice patterns. We utilize a state-of-the-art multilingual TTS model [307] that enables automatic voice generation across diverse datasets in different languages.

The synthesis of pathological voices presents unique challenges, as generated samples must accurately capture the nuanced characteristics specific to each condition. To validate our approach, we carefully assess the generalization capability of synthesized voices by training models exclusively on synthetic data and evaluating their performance on real recordings. This validation process, detailed in Section §6.3.3, ensures that synthetic samples maintain the discriminative features necessary for pathology detection.

To further enhance data diversity, we implement a comprehensive augmentation pipeline. This pipeline includes pitch shifting to simulate variations in vocal range, time stretching to account for different speaking rates, and noise addition to improve robustness to recording conditions. The combination of synthetic data generation and targeted augmentation creates a rich training set that better represents the full spectrum of both healthy and pathological voice characteristics.

Despite these advancements, we must acknowledge the inherent limitations of using synthetic and augmented data for clinical tasks. Real pathological speech, particularly in dysarthria, contains complex neuromuscular irregularities and erratic breath patterns that current TTS models may not fully replicate. These fine-grained physiological nuances represent a significant data bottleneck for synthetic generation.

Consequently, while synthetic data helps mitigate data scarcity, the gap between artificial samples and organic clinical speech remains a challenge for full-scale clinical deployment.

Mixture of Experts framework. We acknowledge that different types of vocal recordings may capture distinct aspects of voice pathologies. Rather than attempting to process all recording types through a single model, we train specialized models for each input type. This specialization allows each model to develop expertise in analyzing specific vocal tasks while maintaining sensitivity to pathology-related features.

To combine insights from these specialized models, we introduce a "shallow" Mixture of Experts framework, which aligns predictions from different models by selecting outputs based on confidence levels. We estimate model confidence through the entropy of predicted probabilities, where lower entropy indicates higher certainty in the prediction. This selection mechanism ensures that final predictions emphasize the most reliable assessments for each input sample.

A key innovation in our approach lies in the specialized pre-training strategy for different recording types. For models analyzing sustained vowels, we utilize pre-training on the AudioSet dataset[1] [237], which contains diverse vocal and non-verbal acoustic events. Models processing sentence readings are pre-trained on LibriSpeech [59], which provides rich exposure to continuous speech patterns. This targeted pre-training strategy leverages the distinct strengths of each dataset: LibriSpeech excels at capturing nuances in continuous speech, whils the diverse acoustic content in AudioSet proves particularly valuable for analyzing sustained vowels, aligning with previous findings [351]. In our experiments, we refer to this specialized configuration as `MoE*`.

6.3.2 Experimental Setup

To evaluate our approach, we conduct experiments across multiple languages and pathological conditions. This section details our experimental protocol, including dataset characteristics, model configurations, and training procedures.[2]

[1] We introduced these models as part of the ARCH framework, discussed in Section §4.5.
[2] https://github.com/koudounasalkis/AI4Voice

Datasets and data processing. We utilize three distinct datasets spanning different languages and recording conditions. Two publicly available collections, the German SVD [364] and Portuguese AVFAD [365], provide broad coverage of voice pathologies. We complement these with an in-house Italian Pathological Voice dataset (IPV) that enables evaluation across additional linguistic contexts.

The *Saarbruecken Voice Database* (SVD) represents our primary German language resource. This collection includes both voice recordings and electroglottography (EGG) data, with each recording session comprising 13 files. These files capture vowels /a, i, u/ across multiple pitch variations (normal, high, low, rising-falling) and include sentence reading tasks. For consistency across our evaluation, we focus specifically on sentence readings and normal-pitch vowel recordings. The dataset provides fine-grained pathology labels without predetermined macro-classes; we concentrate on the six most frequent pathological conditions for classification experiments.

The *Advanced Voice Function Assessment Database* (AVFAD) provides our Portuguese language samples. This collection captures a wide range of vocal tasks, including sustained vowels /a, e, o/, six sentence readings, phonetically balanced text readings, and spontaneous speech. Each task is repeated three times per participant. To maintain consistent evaluation conditions, we concatenate the six sentences for each repetition. Similarly, we combine sustained vowel recordings by repetition, producing three consolidated audio files per vowel.

The *Italian Pathological Voice* (IPV) dataset represents a new contribution to the field. We collected this data through collaboration with multiple Italian phoniatric practices and hospitals. The participant pool includes both healthy individuals seeking routine evaluations and those with various degrees of dysphonia. Each participant underwent comprehensive clinical assessment, including videolaryngostroboscopic examinations and perceptual voice evaluations. The recording protocol focused on two key tasks: sustained production of the vowel /a/ and readings of five phonetically balanced sentences adapted from CAPE-V [366] for Italian. All recordings followed strict standardization procedures, maintaining consistent 30cm mouth-to-microphone distance and ensuring high signal-to-noise ratios above 30.

For all datasets, we apply consistent preprocessing to enable fair comparison. This includes resampling all audio to 16kHz and normalizing amplitude levels. Table 6.1

[3] As macro-classes are not given in SVD, we considered the six most frequent classes.

Table 6.1 **Dataset characteristics.** Overview of the datasets (`Ds`) used in this study, detailing language (`L`), number of healthy (`#H`) and pathological (`#P`) speakers, number of sentence readings (`#S`) and sustained vowel samples (`#V`), number of pathological classes (`#C`), macro classes (`#MC`), and average audio duration in seconds (`T(s)`).

Ds	**L**	**#H**	**#P**	**#S**	**# V**	**#C**	**#MC**	**T (s)**
SVD [364]	DE	687	1356	2043	6129	71	(6)[3]	1.73
AVFAD [365]	PT	346	363	1989	1989	25	8	15.86
IPV	IT	173	340	513	513	15	6	12.89

provides a detailed overview of each dataset's characteristics, including speaker counts, recording durations, and pathology distributions.

Model configurations and training protocol. We compare multiple architectural approaches to establish the effectiveness of our method. We implement both traditional and contemporary models to provide comprehensive performance context.

For baseline comparisons, we reproduce the 1D and 2D CNN architectures from previous works [351, 354]. The 2D CNN implementation explores various MFCC configurations, with our reported results using 40 coefficients based on empirical optimization. These baselines provide important reference points for assessing our transformer-based approach.

Our transformer experiments evaluate multiple state-of-the-art architectures. Following recent work in voice pathology detection [6], we assess wav2vec 2.0 [1], HuBERT [2], and WavLM [57] models in their base configurations. For wav2vec 2.0 and HuBERT, we additionally evaluate variants that we pre-trained on AudioSet in previous work [27].

As the datasets come with no predefined splits, we utilize 10-fold cross-validation to ensure robust performance estimates and assess model stability. Our data augmentation strategy applies targeted transformations including pitch shifting, time stretching, and environmental noise simulation. The augmentation intensity varies by recording type: sentence recordings receive more aggressive augmentation to improve generalization, while vowel recordings undergo more conservative modification to preserve diagnostic acoustic features. Specifically, sentence recordings undergo augmentation with 25% probability, including noise addition (SNR between 0-30dB), speed perturbation ($0.75\times$ to $1.25\times$), and pitch shifting (± 4 semitones). Sustained vowel

recordings receive more conservative augmentation with 10% probability to preserve their core characteristics.

Table 6.2 **Voice Disorder Detection.** Mean ±std results from 10-fold cross-validation across all datasets. Metrics include Accuracy, AUC, and F1 Macro. The best-performing results for each dataset are highlighted in **bold**.

Ds	Model	Accuracy	AUC	F1 Macro
SVD	CNN-1D	.746±.041	.705±.041	.722±.041
	CNN-2D	.799±.025	.734±.025	.747±.024
	HuBERT	.862±.040	.844±.041	.842±.038
	Ours	**.909±.006**	**.911±.005**	**.907±.007**
AVFAD	CNN-1D	.712±.028	.719±.029	.711±.029
	CNN-2D	.835±.019	.834±.021	.834±.021
	HuBERT	.872±.015	.877±.015	.871±.014
	Ours	**.927±.004**	**.931±.004**	**.926±.004**
IPV	CNN-1D	.673±.025	.616±.024	.637±.025
	CNN-2D	.788±.021	.721±.021	.737±.021
	HuBERT	.875±.024	.847±.026	.870±.026
	Ours	**.981±.005**	**.983±.006**	**.978±.005**

6.3.3 Results and Discussion

Our experimental evaluation examines both the overall effectiveness of our approach and the specific contributions of each component. We assess performance on two tasks: voice disorder detection and pathology classification.

Voice disorder detection. The primary task of distinguishing between healthy and pathological voices provides a fundamental test of our method's effectiveness. Table 6.2 presents detailed comparisons across all evaluated models and datasets. Our approach demonstrates significant improvements over existing methods across all evaluation metrics. Compared to traditional architectures, our method achieves substantial gains in AUC: improvements of 0.20-0.36 points over 1D-CNN baselines, 0.10-0.26 points over 2D-CNN approaches, and 0.05-0.13 points beyond standard transformer models. These improvements arise from our method's ability to leverage both synthetic data and multiple recording types effectively. The results demon-

strate that our approach successfully addresses the data scarcity challenge while maintaining high detection accuracy.

Pathology classification. Beyond basic disorder detection, we evaluate the ability of our method to distinguish between specific pathological conditions. This multi-label classification task presents a greater challenge, requiring the model to identify subtle differences between various voice disorders. As shown in Table 6.3, our approach shows particularly strong improvements in macro F1 scores, indicating robust performance across all pathology classes regardless of their representation in the training data.

The performance gains are substantial: improvements of 0.48-0.57 in F1 score compared to 1D-CNNs, 0.38-0.52 versus 2D-CNNs, and 0.11-0.15 beyond standard transformer models. We further enhance these results through an initial pre-training stage on the detection task, ensuring no speaker overlap between pre-training and evaluation (`Ours*` in the table). This pre-training provides an additional boost of 0.01-0.05 in F1 score, demonstrating the value of transfer learning in medical voice analysis.

Table 6.3 **Voice Disorder Classification.** Mean ±std results from 10-fold cross-validation across all datasets. Best-performing results in **bold**.

Ds	Model	Accuracy	F1 Macro
SVD	`CNN-1D`	.437±.025	.280±.024
	`CNN-2D`	.539±.021	.348±.023
	`HuBERT`	.771±.022	.712±.020
	`Ours`	.874±.017	.859±.014
	`Ours*`	**.888±.015**	**.868±.016**
AVFAD	`CNN-1D`	.401±.027	.167±.028
	`CNN-2D`	.509±.025	.266±.024
	`HuBERT`	.693±.028	.538±.026
	`Ours`	.782±.024	.648±.023
	`Ours*`	**.808±.021**	**.703±.020**
IPV	`CNN-1D`	.419±.022	.278±.024
	`CNN-2D`	.521±.019	.335±.021
	`HuBERT`	.764±.025	.710±.023
	`Ours`	.867±.010	.854±.007
	`Ours*`	**.883±.007**	**.871±.006**

Component analysis. To understand the contribution of each element in our pipeline, we conducted detailed ablation studies across all datasets. Table 6.4 presents the impact of progressively adding components to the base transformer model. Data augmentation and synthetic data generation show varying benefits across datasets, with AUC improvements ranging from 0.01 on SVD to 0.10 on IPV. The introduction of our ensemble model with specialized pre-training consistently improves performance

by approximately 0.04 AUC across all configurations. Importantly, attempting to process all data types through a single model (ALL row in the Table) degrades performance below baseline levels. This finding confirms our hypothesis that different recording types provide complementary information best captured through specialized models.

Table 6.4 **Voice disorder detection, ablation study.** Impact of each component on AUC. Best results per model in **bold**, overall best highlighted in light-blue.

Ds	Approach	HuBERT	wav2vec 2.0	WavLM
SVD	base	.844±.041	.842±.038	.842±.035
	+ data++	.851±.032	.849±.036	.849±.032
	+ TTS	.871±.051	.855±.032	.859±.039
	+ MoE	.903±.012	.888±.013	**.884±.019**
	+ MoE*	**.911±.005**	**.894±.011**	-
	ALL	.791±.031	.787±.029	.784±.038
AVFAD	base	.877±.015	.875±.018	.872±.016
	+ data++	.889±.019	.879±.017	.881±.014
	+ TTS	.894±.015	.882±.015	.885±.019
	+ MoE	.917±.007	.902±.014	**.908±.007**
	+ MoE*	**.931±.004**	**.921±.009**	-
	ALL	.865±.012	.861±.017	.863±.013
IPV	base	.847±.026	.874±.021	.832±.029
	+ data++	.903±.016	.894±.019	.845±.022
	+ TTS	.948±.021	.933±.021	.929±.020
	+ MoE	.977±.013	.943±.015	**.930±.012**
	+ MoE*	**.983±.006**	**.970±.005**	-
	ALL	.831±.015	.818±.018	.787±.016

Synthetic data validation. A critical question for our approach concerns the quality of synthetic data generation. To assess this, we conducted experiments training models exclusively on synthetic data and evaluating them on real recordings. Table 6.5 presents these results for the IPV dataset. The performance degradation when using only synthetic training data remains surprisingly modest. Accuracy drops range from 0.07 to 0.14 for detection and 0.05 to 0.12 for classification, across both CNN and transformer architectures. These relatively small decreases indicate that our synthetic data successfully captures the essential characteristics of pathological voice patterns.

Table 6.5 **Voice disorder detection, synthetic data only.** Mean ±std over 10-fold CV on the IPV dataset. Best results are in **bold**, with differences from real data shown in brackets.

Task	Model	Accuracy	F1 Macro
Detection	CNN-1D	.573±.029(−.100)	.564±.030(−.073)
	CNN-2D	.659±.031(−.129)	.592±.033(−.145)
	HuBERT	.794±.028(−.081)	.782±.025(−.088)
	Ours	**.868±.012**(−.113)	**.854±.011**(−.124)
Classification	CNN-1D	.308±.034(−.111)	.221±.036(−.057)
	CNN-2D	.387±.033(−.234)	.284±.032(−.051)
	HuBERT	.651±.027(−.113)	.622±.026(−.088)
	Ours	.749±.025(−.118)	.743±.026(−.111)
	Ours*	**.755±.021**(−.128)	**.751±.022**(−.120)

The models maintain performance well above random chance, approaching their performance on real data.

6.3.4 Summary and Practical Implications

Our data-centric approach to voice pathology detection demonstrates the potential of combining synthetic data generation with specialized model training. The significant improvements in both detection and classification tasks show that transformer models can be effectively adapted for medical applications despite data limitations. The modest performance degradation when training on synthetic data alone validates our approach to addressing data scarcity in medical domains.

The practical applications of this work extend across multiple clinical contexts. For routine medical screening, our system enables more accessible and non-invasive initial assessments, potentially leading to earlier detection of voice disorders. In clinical practice, the ability to analyze both sustained vowels and continuous speech provides physicians with complementary diagnostic information that aligns with traditional examination methods. The robust performance across different languages makes it particularly valuable for deployment in diverse healthcare settings, while its ability to maintain accuracy even with synthetic training data suggests promising applications in medical education and training.

However, several important limitations remain. While our Mixture of Experts framework provides a way to combine information from different recording types, it does not fully exploit the potential relationships between these sources. The approach treats each recording type independently, missing opportunities to model how pathological conditions manifest across different vocal tasks simultaneously. This limitation suggests the need for more sophisticated fusion strategies that can capture the intricate relationships between sustained vowels and continuous speech patterns.

Additionally, our current approach to synthetic data generation, while effective, focuses primarily on expanding dataset size rather than explicitly modeling the relationship between different vocal tasks. A more comprehensive approach would consider how pathological conditions affect different aspects of voice production in a coordinated way. These limitations point toward the need for true multi-source fusion approaches that can more deeply integrate information across different types of vocal recordings while maintaining the benefits of our data-centric improvements.

6.4 Multi-Source Fusion for Voice Pathology Detection

While our previous approach demonstrated the value of combining different vocal tasks through a mixture of experts framework, it treated each recording type independently. This section presents MVP (**M**ulti-source **V**oice **P**athology detection), a more sophisticated approach that explicitly models the relationships between sustained vowels and continuous speech [34]. By developing specialized fusion strategies at multiple levels of abstraction, MVP enables more comprehensive analysis of how voice pathologies manifest across different vocal tasks.

Clinical voice assessment has long recognized that different vocal tasks provide complementary diagnostic information. Sustained vowels offer controlled conditions for analyzing voice quality but miss the dynamics of natural speech. Conversely, continuous speech captures everyday voice use patterns but introduces additional complexity through linguistic content and prosody. Voice pathologies often manifest differently across these tasks: vocal nodules may be more apparent during sustained phonation, while muscle tension disorders might be more evident in continuous

speech [358, 359]. This variation suggests the need for analysis methods that can explicitly model cross-task relationships.

Previous approaches to automated voice analysis have typically focused on single recording types in isolation. Even methods that consider multiple sources, like our earlier work, often process them independently before combining predictions. This limitation fails to capture the potential interactions between different vocal tasks that might provide key diagnostic information. MVP addresses this gap by introducing three distinct fusion strategies: waveform concatenation, intermediate feature fusion, and decision-level combination. Each strategy represents a different approach to integrating information across recording types, enabling us to identify the most effective level for cross-source learning.

6.4.1 Methodology

MVP implements a three-stage architecture designed to process and combine information from multiple recording sources effectively. The first stage uses source-specific backbone models to extract relevant features from each recording type. The second stage applies one of several fusion strategies to combine information across sources. The final stage makes diagnostic decisions based on the fused representations. An overview of the entire framework is shown in Figure 6.3.

Feature extraction and backbone models. Let X_{SV} denote a sustained vowel recording and X_{S} denote a sentence recording. The feature extraction process differs depending on the chosen fusion strategy. For waveform concatenation, a single backbone processes the combined input. For intermediate fusion and decision-level combination, each input type is processed by its specialized backbone model. The sentence recordings are processed by HuBERT pre-trained on LibriSpeech [59], optimizing for linguistic and prosodic feature extraction. Sustained vowel recordings use HuBERT pre-trained on AudioSet [237], focusing on vocal quality and non-semantic acoustic features. These backbones generate frame-level representations according to:

$$H_{\mathrm{SV}} = \mathrm{HuBERT}_{\mathrm{AS}}(X_{\mathrm{SV}}), \quad H_{\mathrm{S}} = \mathrm{HuBERT}_{\mathrm{LS}}(X_{\mathrm{S}}) \tag{6.1}$$

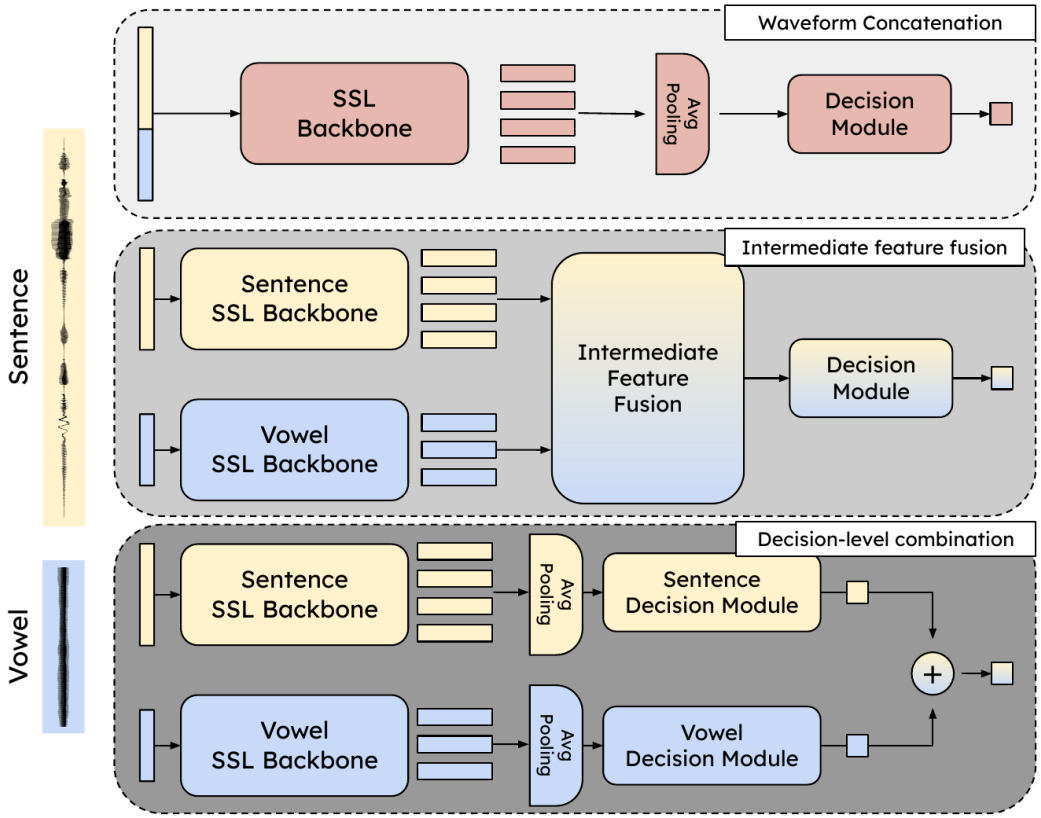


Fig. 6.3 **Overview of the MVP framework.** Illustration of the three main stages: waveform-level concatenation (top), intermediate feature fusion (middle), and decision-level integration (bottom).

where $H_{\mathrm{SV}} \in \mathbb{R}^{T_{\mathrm{SV}} \times d}$ and $H_{\mathrm{S}} \in \mathbb{R}^{T_{\mathrm{S}} \times d}$ contain temporal information from each recording type, T_{SV} and T_{S} represent the respective sequence lengths, and d is the latent dimensionality of the representations.

Fusion strategies. We investigate three distinct approaches to combining information from sustained vowel and sentence recordings, each operating at a different level of abstraction in the processing pipeline.

Waveform Concatenation (WC). The most direct approach (Figure 6.3, top) concatenates the raw audio signals before processing, preserving all original information but potentially introducing artifacts due to the different characteristics of each source type. This method, while simple, allows the model to learn cross-source relationships from the earliest processing stages.

Intermediate Feature Fusion (IFF). This strategy (Figure 6.3, middle) combines features extracted by the specialized backbones through various fusion methods detailed in the next paragraph. It maintains source-specific characteristics while enabling explicit cross-source learning at the representation level.

Decision-Level Combination (DLC). This approach (Figure 6.3, bottom) extends our previous work [33] by implementing an ensemble that dynamically selects between source-specific predictions through probability averaging of individual backbone outputs.

Intermediate Feature Fusion methods. Given the feature sequences $H_{\text{SV}} \in \mathbb{R}^{T_{\text{SV}} \times d}$ and $H_{\text{S}} \in \mathbb{R}^{T_{\text{S}} \times d}$, we explore five distinct methods for generating a fused representation $Z_{\text{fused}} \in \mathbb{R}^d$, each capturing different aspects of cross-source interactions [367].

Simple concatenation. Our baseline fusion approach first applies average pooling to each sequence independently:

$$Z_{\text{Concat}} = [Z_{\text{SV}}^{\text{Avg}}; Z_{\text{S}}^{\text{Avg}}] \in \mathbb{R}^{2d} \tag{6.2}$$

where $Z_{\text{SV}}^{\text{Avg}}$ and $Z_{\text{S}}^{\text{Avg}}$ represent global representations for each source. This method preserves source-specific global characteristics while providing a strong baseline for comparison.

Attention pooling. This method concatenates sequences along the time dimension and applies learned attention weights to capture the relative importance of different time steps. Using a learnable vector $w \in \mathbb{R}^d$, attention scores are computed as:

$$\alpha_t = \frac{\exp(w^\top H_t)}{\sum_{t'} \exp(w^\top H_{t'})}, \quad Z_{\text{AP}} = \sum_t \alpha_t H_t \in \mathbb{R}^d \tag{6.3}$$

where H_t represents features at time step t from the concatenated sequence $[H_{\text{SV}}; H_{\text{S}}]$.

Gating mechanism. Following [368], we implement adaptive weighting between sources through a learned gating mechanism:

$$G = \sigma(W[Z_{\text{SV}}^{\text{AP}}; Z_{\text{S}}^{\text{AP}}]), \quad Z_{\text{Gating}} = G \odot [Z_{\text{SV}}^{\text{AP}}; Z_{\text{S}}^{\text{AP}}] \tag{6.4}$$

where W is a learnable matrix and σ represents the sigmoid function. This mechanism enables dynamic adjustment of the contribution of each source.

Feature-wise Linear Modulation (FiLM). FiLM [369] enables bidirectional cross-source interactions through mutual modulation:

$$Z_{\mathrm{S}}^{\mathrm{FiLM}} = Z_{\mathrm{S}}^{\mathrm{AP}} \odot (1 + W_{\gamma}^{s} Z_{\mathrm{SV}}^{\mathrm{AP}}) + W_{\beta}^{s} Z_{\mathrm{SV}}^{\mathrm{AP}} \tag{6.5}$$

$$Z_{\mathrm{SV}}^{\mathrm{FiLM}} = Z_{\mathrm{SV}}^{\mathrm{AP}} \odot (1 + W_{\gamma}^{v} Z_{\mathrm{S}}^{\mathrm{AP}}) + W_{\beta}^{v} Z_{\mathrm{S}}^{\mathrm{AP}} \tag{6.6}$$

where $W_{\gamma}^{s}, W_{\gamma}^{v}, W_{\beta}^{s}, W_{\beta}^{v}$ are learnable parameters for scale and shift operations. The final representation combines both modulated features:

$$Z_{\mathrm{FiLM}} = [Z_{\mathrm{S}}^{\mathrm{FiLM}}; Z_{\mathrm{SV}}^{\mathrm{FiLM}}] \tag{6.7}$$

Transformer encoder. This method first concatenates sequences temporally:

$$H_{\mathrm{combined}} = [H_{\mathrm{SV}}; H_{\mathrm{S}}] \in \mathbb{R}^{(T_{\mathrm{SV}} + T_{\mathrm{S}}) \times d} \tag{6.8}$$

The combined sequence is processed through L transformer encoder layers, enabling fine-grained interaction between time steps from both sources through self-attention mechanisms. The final representation is obtained via attention pooling:

$$Z_{\mathrm{TE}} = \mathrm{AP}(\mathrm{TE}_{L}(H_{\mathrm{combined}})) \in \mathbb{R}^{d} \tag{6.9}$$

Decision module. The final stage of our pipeline transforms the fused representation Z into a diagnostic prediction through a fully connected (FC) layer followed by sigmoid activation:

$$\hat{y} = \mathrm{Sigmoid}(\mathrm{FC}(Z)) \tag{6.10}$$

where $\hat{y} \in [0, 1]$ represents the probability of voice pathology presence.

6.4.2 Experimental Setup

Building on our previous experimental framework, this section details the specific setup for evaluating the MVP approach. We focus on highlighting the key differences in dataset preparation, model configurations, and evaluation protocols that are unique to our multi-source fusion strategies.[4]

[4] https://github.com/koudounasalkis/MVP

Dataset preparation. While we utilize the same three datasets (SVD, AVFAD, and IPV) described in Section 6.3.2, our preparation emphasizes the multi-source nature of MVP. For SVD, we pair normal pitch vowel recordings with sentence readings for each subject, maintaining the dataset's balance of 687 healthy and 1,356 pathological voices. In AVFAD, we create consistent paired inputs by randomly selecting one of the six available sentence recordings and one sustained vowel (`/a/`, `/e/`, or `/o/`) for each of the 709 subjects. For IPV, we pair the single sustained `/a/` vowel recording with a randomly selected sentence from the five available for each of the 513 subjects. This pairing process ensures that each subject has a consistent set of multi-source inputs, enabling our fusion strategies to leverage both sustained vowel and continuous speech information.

Model configurations. We compare several model configurations to assess the effectiveness of different fusion strategies. For single-source baselines, we implement HuBERT models pre-trained on either LibriSpeech (`LS`) or AudioSet (`AS`), processing only one recording type. An additional baseline uses HuBERT (`LS`) trained on a mix of sentences and vowels.
The waveform concatenation (`WC`) approach uses a single HuBERT backbone to process the concatenated raw audio from both sources. For intermediate feature fusion (`IFF`), we evaluate both frozen (17.98M trainable parameters) and fine-tuned (206.73M parameters) backbones, extracting representations from the 5th layer for fusion. The decision-level combination (`DLC`) strategy trains separate models for each source (`LS` for sentences, `AS` for vowels) or implements a combined approach using both specialized backbones.
All transformer-based models use HuBERT in its base configuration (94.64M parameters). For the Transformer Encoder (`TE`) fusion method, we use 2 encoder layers based on empirical performance.

Training and evaluation protocol. We maintain a 10-fold cross-validation approach with speaker-independent splits as in [33]. Audio preprocessing includes $16kHz$ resampling with zero-mean and unit variance normalization. Further details are provided in the official project repository.

Data augmentation follows differentiated strategies for different recording types, as detailed in Section 6.3.2. This approach maintains signal integrity while ensuring appropriate data diversity for each recording type.

Performance evaluation uses accuracy, macro F1 score, and AUC-ROC, averaged across folds with standard deviations.

Table 6.6 **MVP, model performance**. Mean±std performance of different fusion strategies across three datasets. Best results are shown in **bold**, and second-best are underlined. Models are pre-trained on either LibriSpeech (`LS`) or AudioSet (`AS`). →`Sent` denotes fine-tuning on read sentences, →`Vowel` on sustained vowels, and →`Mix` on both. A ❄ symbol marks frozen backbones. Intermediate Feature Fusion (IFF) is implemented using a Transformer Encoder (TE).

Model	# Params	SVD			AVFAD			IPV		
		Acc.	F1	AUC	Acc.	F1	AUC	Acc.	F1	AUC
Single-Source Baselines										
LS→Sent	94.64M	.873±.058	.849±.062	.850±.048	.872±.015	.871±.014	.877±.015	.875±.024	.870±.026	.847±.026
LS→Vowel	94.64M	.747±.075	.724±.074	.732±.084	.714±.051	.705±.061	.714±.061	.622±.064	.617±.062	.620±.062
AS→Sent	94.64M	.817±.060	.801±.061	.810±.052	.852±.045	.850±.047	.855±.050	.828±.039	.823±.038	.815±.044
AS→Vowel	94.64M	.779±.075	.747±.068	.760±.079	.798±.043	.758±.056	.756±.058	.676±.059	.649±.060	.665±.055
LS→Mix	94.64M	.780±.028	.791±.031	.765±.025	.827±.017	.826±.019	.827±.019	.840±.018	.831±.015	.831±.016
Waveform Concatenation (WC)										
LS	94.64M	.896±.054	.882±.053	.891±.056	.907±.054	.906±.055	.908±.052	.888±.066	.881±.067	.886±.060
AS	94.64M	.875±.063	.869±.061	.873±.068	.889±.042	.884±.045	.885±.049	.836±.031	.832±.032	.831±.037
Intermediate Feature Fusion (IFF)										
LS+AS ❄	17.98M	.826±.036	.809±.035	.832±.023	.833±.032	.831±.032	.834±.033	.813±.048	.806±.045	.809±.048
LS+AS	206.73M	**.958±.063**	**.953±.067**	**.958±.062**	**.962±.040**	**.962±.039**	**.963±.038**	**.939±.044**	**.931±.054**	**.936±.053**
Decision-Level Combination (DLC)										
LS+LS	189.28M	.885±.062	.863±.070	.860±.070	.881±.095	.874±.111	.879±.099	.882±.083	.873±.084	.862±.064
AS+AS	189.28M	.864±.072	.855±.076	.857±.075	.872±.034	.863±.034	.866±.035	.837±.061	.826±.095	.827±.102
LS+AS	189.28M	.898±.051	.884±.058	.896±.058	.888±.092	.887±.108	.889±.096	.896±.072	.877±.074	.882±.062

6.4.3 Results and Discussion

Our experimental evaluation examines both the overall effectiveness of our multi-source approach and the specific contributions of different fusion strategies. We assess performance on voice disorder detection across multiple datasets and languages, comparing our method against single-source baselines and analyzing the impact of various architectural choices.

Overall system performance. Table 6.6 presents a comprehensive comparison of different system configurations across our three evaluation datasets. The results demonstrate that our multi-source approach consistently outperforms single-source baselines across all datasets. The `IFF-TE` method with fine-tuned backbones achieves the highest AUC scores: 95.8% (SVD), 96.3% (AVFAD), and 93.6% (IPV). This

represents a 10-13% improvement over the best single-source baseline, clearly demonstrating the advantage of our multi-source approach.

When examining single-source baselines, models consistently perform better on sentence readings compared to sustained vowels. This suggests that sentence readings may contain richer diagnostic information, possibly capturing both phonation quality and dynamic speech characteristics. However, training a single model on a mixture of sources (`LS→Mix`) proves ineffective, likely due to the model's inability to adapt to the diversity of recording types.

The significant performance improvement observed with our multi-source approach highlights the complementary value of sustained vowels when combined effectively with sentence readings. Intermediate Feature Fusion (`IFF`) also outperforms other fusion strategies such as Waveform Concatenation (`WC`) and Decision-Level Combination (`DLC`). Interestingly, `WC` performs better than `DLC` in two out of three datasets while requiring only a single model, effectively halving the number of parameters. This suggests that even in simpler settings, concurrent access to both sources allows `WC` to learn cross-source patterns, supporting the value of joint analysis.

Analysis of fusion strategies. Table 6.7 compares different `IFF` strategies across all datasets. The Transformer Encoder (`TE`) consistently achieves the best overall performance, particularly on SVD and AVFAD datasets. While FiLM and Attention Pooling provide strong alternatives, the significant performance gap between learned fusion strategies and simple concatenation (up to 4.7% on AVFAD) highlights the importance of modeling fine-grained interactions between sources. These results suggest that effective voice pathology detection requires careful modeling of how different vocal characteristics manifest across both sustained and dynamic speech patterns.

Table 6.7 **IFF fusion strategies comparison.** AUC scores are reported, with best results in **bold** and second-best underlined.

Method	SVD	AVFAD	IPV
`Concat`	.918±.060	.920±.044	.915±.050
`AP`	.948±.060	.955±.040	.929±.047
`TE`	**.958±.062**	**.963±.038**	**.936±.053**
`Gating`	.947±.068	.956±.045	.926±.049
`FiLM`	.951±.062	.961±.041	.934±.053

Feature extraction depth. Table 6.8 demonstrates the impact of feature extraction depth on model performance. While performance remains relatively stable across different layers, the 5th layer consistently provides optimal results. This

suggests that mid-level representations achieve the best balance between preserving source-specific characteristics and enabling effective cross-source integration. A learned weighted combination across all layers shows competitive but not superior performance, indicating that optimal weighting might require additional training data.

Table 6.8 **Feature extraction layer depth.** AUC scores are reported, with best results in **bold** and second-best underlined.

Layer	SVD	AVFAD	IPV
4th	.944±.054	.945±.043	.924±.048
5th	**.958±.062**	**.963±.038**	**.936±.053**
6th	.950±.055	.943±.046	.923±.060
7th	.948±.066	.954±.042	.929±.061
Last	.954±.061	.958±.037	.932±.072
Weighted	.953±.059	.957±.042	.932±.057

Backbone models. Table 6.9 reveals the impact of different backbone configurations on system performance. The combination of HuBERT models pre-trained on LibriSpeech (LS) for sentences and AudioSet (AS) for sustained vowels consistently outperforms other configurations across all datasets. This aligns with the distinct strengths of each dataset: LibriSpeech excels at capturing nuances in sentence readings, while AudioSet's diverse acoustic content proves particularly valuable for analyzing sustained vowels. To process sustained vowels, we also evaluated voc2vec [28], which we presented in Chapter 5. Although it does not outperform the HuBERT LS-AS combination, its solid performance underscores the benefits of domain-specific pre-training for vocal analysis.

The consistent superiority of the LS-AS combination demonstrates that acoustic event coverage in AudioSet provides more effective representations for capturing pathological voice characteristics. These findings highlight the importance of matching pre-training data to the specific requirements of different vocal tasks. The success of LibriSpeech pre-training for sentence analysis and AudioSet pre-training for sustained vowels suggests that each recording type benefits from exposure to different types of acoustic patterns during pre-training.

Table 6.9 **Backbone models.** AUC scores are reported. H=HuBERT, v2v=voc2vec, LS=LibriSpeech, AS=AudioSet. Best results are in **bold**, and second-best are underlined.

Sent.	Vowel	SVD	AVFAD	IPV
H-LS	H-LS	.942±.059	.951±.042	.917±.067
H-AS	H-AS	.935±.072	.939±.042	.901±.048
H-LS	H-AS	**.958±.062**	**.963±.038**	**.936±.053**
H-LS	v2v	.953±.056	.958±.044	.929±.044

6.4.4 Summary and Practical Implications

Our experiments demonstrates the clear advantages of multi-source analysis for voice pathology detection. The consistent improvements across datasets, languages, and pathology types validate our approach to combining information from different vocal tasks. The success of intermediate feature fusion, particularly with transformer-based integration, suggests that the relationship between sustained and continuous vocal productions contains important diagnostic information.

Several practical considerations emerge from our analysis. While our approach requires more computational resources than single-source methods, the substantial performance gains justify this increase. The ablation studies provide clear guidance for implementing these methods in practical settings, particularly regarding the choice of feature extraction layer and fusion strategy. These findings have implications beyond pathology detection, potentially informing both clinical practice and our understanding of voice disorders.

Looking forward, our work establishes important foundations for next-generation voice analysis systems. The success of transformer-based fusion strategies suggests promising directions for integrating multiple diagnostic inputs in other medical applications. While computational requirements present deployment challenges, the clear performance benefits justify continued development of efficient multi-source analysis methods. As healthcare increasingly moves toward comprehensive digital assessment tools, our approach provides a blueprint for developing more sophisticated and reliable diagnostic systems.

6.5 Generative Error Correction for Dysarthric Speech Recognition

While previous sections have explored voice pathology detection and multi-source analysis, a particularly challenging aspect of medical speech technology involves accurate transcription of dysarthric speech. Despite remarkable advances in general-purpose speech recognition, individuals with dysarthria face significant barriers when using conventional ASR systems. These challenges arise from the distinctive

characteristics of dysarthric speech, including irregular articulation, atypical prosody, and inconsistent speaking rates [5, 348].

This technological gap has implications that surpass simple usability concerns. As voice interfaces become increasingly central to daily life, from digital assistants to accessibility tools, the inability to accurately process dysarthric speech risks excluding individuals with motor speech disorders from these technological advances. Although modern ASR systems benefit from self-supervised learning and large-scale training, their error rates on dysarthric speech often exceed 30%, limiting their practicality in real-world applications [333, 348].

This section presents a novel two-stage approach that combines state-of-the-art ASR with large language model-based generative error correction (GER) [35]. Our method builds on a key insight: while ASR systems may struggle to produce correct transcriptions for dysarthric speech, they often capture useful acoustic information in their N-best hypotheses. By leveraging a large language model to analyze these hypotheses collectively, we can improve transcription accuracy while maintaining linguistic coherence.

Our investigation addresses two fundamental questions about dysarthric speech recognition. First, we examine whether general-purpose ASR models possess sufficient acoustic modeling capacity to capture dysarthric speech patterns, even when their top predictions are incorrect. Second, we explore how large language models can leverage their linguistic knowledge to identify correct transcriptions by analyzing patterns across multiple ASR hypotheses. Through comprehensive experiments on the Speech Accessibility Project dataset [348], we demonstrate the effectiveness of our approach while highlighting specific challenges in different speaking contexts.

6.5.1 Methodology

Our approach investigates two fundamental aspects of dysarthric speech recognition: the ability of ASR systems to capture accurate information in their N-best hypotheses, and the potential for language models to effectively distill this information. We develop a two-stage framework combining ASR with generative error correction (GER), separating acoustic processing from linguistic error correction to enable detailed analysis of each component's contribution. Figure 6.4 shows an overview of the proposed framework.

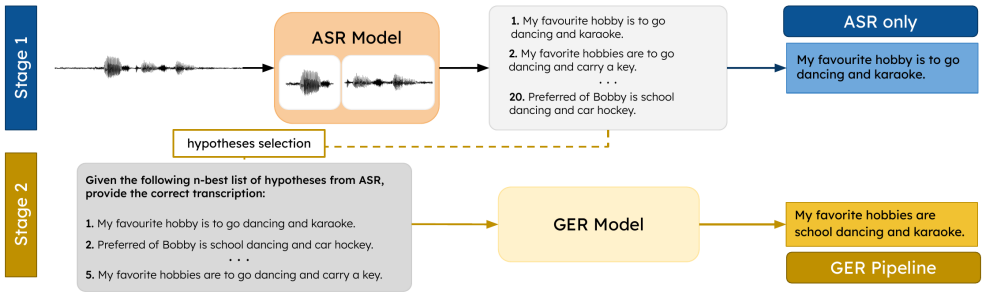


Fig. 6.4 **Two-stage dysarthric speech recognition framework, overview.** Stage 1 generates a 20-best list of hypotheses from the input audio using the ASR model. Stage 2 selects diverse candidates and applies the GER model to analyze them jointly, producing a refined final transcription.

Speech recognition stage. The first stage of our pipeline employs Whisper [13] for acoustic processing, implemented within the WhisperX [329] framework to enable efficient processing of long-form audio. We explore two distinct approaches to handling dysarthric speech. The zero-shot evaluation tests the inherent capability of general-purpose ASR models to process dysarthric speech without adaptation. The fine-tuning approach allows the model to specialize specifically for dysarthric speech patterns.

For each audio input, our system generates N transcription hypotheses through beam search. This N-best generation strategy enables the capture of different interpretations for unclear or challenging speech segments. For recordings exceeding 30 seconds, WhisperX manages segmentation through integrated voice activity detection. To maintain consistency in these multi-segment recordings, we concatenate corresponding ranked hypotheses across segments, ensuring that all top-1 hypotheses form a complete top-1 transcription.

Hypothesis selection strategy. To balance diversity with computational efficiency, we implement a selection algorithm that identifies 5 maximally informative hypotheses from an initial pool of 20. The selection process follows a three-step procedure. First, we retain the top-scoring hypothesis to preserve the highest confidence transcription. Second, we compute normalized edit distances between all remaining hypotheses to measure their dissimilarity. Third, we iteratively select hypotheses that maximize the minimum distance to previously selected ones. This approach ensures the selected hypotheses represent genuinely different interpretations rather than minor variations of the same transcription.

Given the following n-best list of hypotheses from ASR, provide the correct transcription:

- Hypothesis 1
- Hypothesis 2
- ...
- Hypothesis 5

Fig. 6.5 **GER Prompt.** Template of model prompt.

Generative error correction. The second stage of our framework employs a sequence-to-sequence model to analyze the selected hypotheses collectively and generate refined transcriptions. We utilize the FlanT5 [370] model, which has demonstrated strong capabilities in error correction tasks [363]. The model processes variable-length input sequences through a structured prompt, generating coherent text while maintaining semantic meaning. An example of structured prompt received as input is shown in Figure 6.5.

To maintain computational efficiency while specializing in transcription correction, we train the GER model using LoRA adaptation. This approach introduces approximately 1% additional trainable parameters while keeping the base model frozen. The GER stage serves three important functions. First, it analyzes patterns across multiple hypotheses to identify consistent elements, leveraging the collective information present in different ASR interpretations. Second, it employs linguistic knowledge to correct grammatical errors and resolve ambiguities, ensuring outputs maintain natural language structure. Third, it produces final transcriptions that balance acoustic evidence with linguistic plausibility, helping to avoid common ASR errors while maintaining fidelity to the original speech.

Training protocol. Our framework implements a two-phase training approach to ensure each component develops specialized capabilities while maintaining complementary functionality. The first phase focuses on acoustic modeling, either using the ASR model in zero-shot mode or fine-tuning it specifically for dysarthric speech. The second phase trains the GER model using pairs of N-best hypotheses and ground truth transcriptions from the training set.

This separation of acoustic and linguistic processing allows each component to optimize for its specific task. The ASR model focuses on capturing acoustic patterns and generating diverse hypotheses, while the GER model specializes in analyzing these hypotheses collectively to produce accurate final transcriptions. The approach enables detailed analysis of each component's contribution while maintaining practical computational efficiency.

6.5.2 Experimental Setup

Our evaluation employs the Speech Accessibility Project Challenge (SAPC) dataset [348], comprising 105.76 hours of dysarthric speech for training and 39.56 hours for development. The dataset encompasses multiple speaking contexts: Digital Assistant Commands (DAC), Sentences from Novels (SN), Spontaneous Speech (SS), and single-word utterances (SW). This diversity enables comprehensive assessment across different speaking scenarios.

Dataset preparation. To enhance model robustness, we augment the SAPC dataset with additional speech sources. The TORGO [371] dataset contributes 10.13 hours of dysarthric speech (5.71 hours natural and 4.42 hours synthetic), providing exposure to diverse pathological speech patterns. We also incorporate VoxPopuli [63] (500+ hours of native and non-native English speech) to maintain general speech recognition capabilities while adapting to dysarthric characteristics.

Our data processing pipeline, inspired by CrisperWhisper [372], implements three key augmentation strategies. First, we inject background noise from MUSAN [373] at audio segment boundaries with 50% probability, preventing overfitting to temporal positions while simulating realistic acoustic conditions. Second, we introduce pure noise samples during training (1% probability) to mitigate hallucination artifacts. Third, we apply either time stretching (0.85×-1.15× range) or SpecAugment [374] with 25% probability to address variable speaking rates and enhance spectral robustness.

Transcription protocol. The SAPC dataset provides detailed transcriptions capturing both main speech content and disfluencies. For spontaneous speech, transcriptions include contextual information in square brackets, while prompted speech contains only speaker responses. Speech events with clear acoustic evidence are preserved in parentheses, such as partial word attempts ("*(che- che-) checkout*") or

sound repetitions ("*(d-*) demo*"). Special tags mark contextual events: "*(cs:...)*" for interviewer speech and "*(ss:...)*" for off-prompt speaker utterances. Our processing retains all parenthetical content corresponding to acoustic events while removing square-bracketed contextual information. This approach preserves the relationship between transcriptions and actual acoustic evidence while standardizing the evaluation format.

Model configuration. For the ASR stage, we evaluate two Whisper [13] configurations: `LARGE-V3` in zero-shot mode and `LARGE-V2` with fine-tuning.[5] The choice of `LARGE-V2` for adaptation follows empirical observations of instability in `LARGE-V3` fine-tuning on dysarthric speech, where the latter exhibited degradation through repetitive patterns and incomplete utterances. This suggests that `LARGE-V3`'s pretrained representations, while effective for general ASR, may require specialized adaptation strategies for dysarthric speech.

The GER stage employs FlanT5 [370] in two configurations: XL (3B parameters) and XXL (11B parameters). We implement LoRA adaptation targeting all linear layers, introducing approximately 1% additional trainable parameters while maintaining the base model frozen. This efficient adaptation strategy enables specialized error correction while managing computational requirements.

Evaluation protocol. Performance assessment uses three evaluation sets: development, `TEST-1`, and `TEST-2`. The development set enables model optimization, `TEST-1` provides ongoing feedback through a public leaderboard, and `TEST-2` remains hidden until challenge completion to ensure unbiased final evaluation.[6]

We employ two complementary metrics: WER for literal transcription accuracy and Semantic Score (SemScore) [375] for meaning preservation. SemScore combines BERTScore [202] for semantic similarity, phonetic distance for pronunciation accuracy, and natural language inference for entailment relationships. Following challenge guidelines, each utterance is evaluated against both disfluency-inclusive and clean reference transcripts, with scoring using the reference yielding lower normalized edit distance.

[5] https://github.com/MorenoLaQuatra/GER4Dys

[6] This paper corresponds to a submission to the SAPC challenge at Interspeech 2025.

Table 6.10 **ASR and GER.** Model performance comparison across configurations.

ASR Model	FT	GER	Dev		TEST-1		TEST-2	
			WER	SemScore	WER	SemScore	WER	SemScore
Lv3	✗	✗	11.60	83.91	11.39	84.82	14.49	78.83
Lv3	✗	3B	7.91	89.42	10.99	86.45	13.87	79.88
Lv3	✗	11B	7.34	90.32	**10.63**	86.96	13.64	80.4
Lv2	✓	✗	7.17	91.94	10.90	87.24	13.04	81.89
Lv2	✓	3B	**6.40**	**92.47**	10.67	**87.51**	**12.89**	**82.16**

6.5.3 Results and Discussion

Our experimental analysis examines both the general effectiveness of our two-stage approach and the specific contributions of acoustic and linguistic modeling to dysarthric speech recognition. Through comprehensive evaluation across different speaking contexts and model configurations, we assess both the capabilities and limitations of our system.

Overall system performance. Table 6.10 presents performance comparisons across ASR and GER configurations on development and test sets. The baseline Whisper `LARGE-V3` model achieves 11.60% WER on the development set without adaptation, demonstrating strong zero-shot capabilities for dysarthric speech. However, performance varies considerably across evaluation sets: while achieving 11.39% WER on `TEST-1`, performance degrades to 14.49% WER on `TEST-2`, indicating significant variations in speech complexity across test sets.

Adding generative error correction with the 11B parameter model substantially improves performance, reducing development set WER to 7.34% and `TEST-1` WER to 10.63%. The corresponding improvement in semantic scores (from 83.91 to 90.32 on development set) indicates that GER not only corrects transcription errors but also enhances semantic understanding. Fine-tuning Whisper `LARGE-V2` provides additional substantial gains, achieving 7.17% WER and 91.94 SemScore without GER on the development set, demonstrating effective acoustic model adaptation for dysarthric speech.

The combination of fine-tuning with GER achieves our best overall performance: 6.40% WER and 92.47 SemScore on development, and 12.89% WER on `TEST-2`. These results demonstrate the complementary benefits of acoustic adaptation and

linguistic error correction. However, the consistent performance gap between development and test sets (approximately 6 percentage points in WER) highlights the ongoing challenge of building robust systems for diverse speaking patterns.

Model scaling and computational efficiency. Our experiments with model scaling reveal interesting trade-offs between performance and computational requirements. While scaling the GER model from 3B to 11B parameters improves development set performance, we observe diminishing returns on test sets. This suggests that computational resources might be more effectively invested in other system components. In terms of practical deployment, our system maintains reasonable efficiency despite its sophisticated components. Using a single NVIDIA *A*100 GPU, our fine-tuned Whisper `LARGE-V2` achieves an average inference time of 0.55*s* per sample. The complete pipeline including GER averages 0.69*s* per sample, remaining practical for real-world applications despite the high computational demands of the models.

Analysis of N-best generation. Examining the impact of N-best list size provides important insights into our system's acoustic modeling capabilities. Table 6.11 shows that increasing the number of hypotheses from 1 to 20 provides diminishing returns. Most gains are achieved by the first 5 hypotheses, reducing WER from 12.37% to 11.71%. This pattern supports our first research question, suggesting that the ASR model effectively captures acoustic information within its top hypotheses.

Table 6.11 **N-best list size.** Impact on final performance, WER and SemScore.

ASR Model	N	Dev	
		WER	SemScore
Lv3	1	12.37	83.07
Lv3	5	11.71	83.79
Lv3	10	11.62	83.89
Lv3	20	**11.60**	**83.91**

The modest improvements beyond 5 hypotheses (only 0.11% WER reduction from 5 to 20) indicate that additional candidates often represent minor variations rather than fundamentally different interpretations. Our diversity-based selection algorithm effectively maintains meaningful variation while managing computational costs.

Performance across speaking contexts. Table 6.12 reveals distinct patterns in how our approach affects different types of speech. For Digital Assistant Commands (DAC), which constitute the majority of our dataset, GER reduces WER by 0.86% absolute while improving semantic scores by 0.51 points. Similar improvements appear in Sentences from Novels (SN), with a 0.53% WER reduction and 0.45-point semantic improvement. Spontaneous Speech (SS) shows comparable WER reduction

Table 6.12 **GER, bucket-analysis.** Development set performance broken down by utterance category: DAC (Digital Assistant Commands), SN (Novel Sentences), SS (Spontaneous Speech), SW (Single Words).

Category	**# Samples**	**w/o GER**		**w/ GER**	
		WER ↓	SemScore ↑	WER ↓	SemScore ↑
DAC	15,066	6.47	92.77	$5.61_{-0.86}$	$93.28_{+0.51}$
SN	3,746	5.74	92.96	$5.21_{-0.53}$	$93.41_{+0.45}$
SS	2,313	10.86	87.27	$10.31_{-0.55}$	$88.14_{+0.87}$
SW	130	63.08	49.84	$63.08_{+0.00}$	$49.46_{-0.38}$

Table 6.13 **Spontaneous speech, qualitative example**. Multiple N-best hypotheses shown, with correct words highlighted in green and errors in red.

Reference	My favorite pet is the one that sits on my lap.
ASR Output	My favorite play is the one that's set on Monday.
GER Output	My favorite pet is the one that sits on my lap.
N-best Hypotheses:	
1.	My favorite play is the one that's set on Monday.
2.	My favorite pet is the one that sits on my lap.
3.	My favorite player is the one that's in Orlando.
4.	My favorite play is the ones that sit on the.
5.	My favorite pick is the one that said "Wonder."

(0.55%) but achieves a larger semantic improvement (0.87 points), indicating that GER's corrections are particularly meaningful for conversational speech.

However, our analysis reveals a critical limitation in processing single words (SW). With an extremely high WER of 63.08% and no improvement from GER, this category represents a significant challenge. Detailed error analysis shows that the system consistently misinterprets isolated words as common phrases or expressions. For example, "*football*" becomes "*What law?*", "*tape*" becomes "*Hey*", and "*Birmingham*" is transcribed as "*Lonnie Hill*". This pattern suggests a strong bias towards generating complete utterances rather than single words.

Qualitative analysis. Table 6.13 presents a detailed example demonstrating both the strengths and mechanisms of our approach. In this case from spontaneous speech, our diversity-based hypothesis selection successfully preserves the correct transcription despite it not being the top-ranked hypothesis. While the initial ASR

output contains errors (replacing "*pet*" with "*play*", "*sits*" with "*set*", and "*lap*" with "*Monday*"), the correct interpretation appears as the second hypothesis. The GER model successfully identifies and selects this correct version by analyzing patterns across the diverse candidate set. This example illustrates how our approach can overcome individual ASR errors by leveraging multiple hypotheses. However, it also highlights the system's preference for complete, grammatically coherent utterances, a characteristic that becomes problematic when processing single words.

6.5.4 Summary and Practical Implications

Our investigation through the Speech Accessibility Project Challenge reveals both significant opportunities and remaining challenges in dysarthric speech recognition. The strong performance of zero-shot Whisper (11.60% WER) answers our first research question, demonstrating that general-purpose ASR models can still capture dysarthric speech patterns. The consistent improvements from GER across most speaking styles validates language models' ability to leverage multiple hypotheses, though with diminishing returns at larger scales.

The practical implications of our work are particularly relevant for accessibility technology. For individuals with dysarthria, our system's improved accuracy on continuous speech enables more reliable interaction with voice-controlled devices and digital assistants. The effectiveness of our two-stage approach in maintaining both transcription accuracy and semantic meaning suggests potential applications in medical documentation, educational support, and everyday communication aids. The ability of the system to process spontaneous speech effectively makes it particularly valuable for natural interaction scenarios, while its computational efficiency enables practical real-world deployment.

However, important limitations emerged that require further research attention. The poor performance on isolated words (63.08% WER) suggests current models may overly favor complete utterances, while the significant gap between development and test set performance indicates ongoing challenges in generalizing across different dysarthric speech patterns. These limitations have direct implications for practical applications, particularly in command-and-control interfaces where single-word accuracy is crucial.

Looking forward, these findings suggest several promising directions for both research and application. The success of our generative error correction approach could be extended to other forms of impaired speech, potentially benefiting individuals with different types of speech disorders. While specialized architectures for isolated word recognition remain a fundamental challenge, our work establishes important foundations for more inclusive speech technology. As voice interfaces become increasingly prevalent in daily life, continued advancement in dysarthric speech recognition will be essential for ensuring equal access to these technologies.

6.6 Conclusions

This chapter has presented three complementary approaches to advancing medical speech technology, each addressing distinct challenges in the application of foundation models to clinical contexts. Our investigations span voice pathology detection, multi-source analysis, and dysarthric speech recognition, demonstrating both the potential and current limitations of these technologies in medical applications.

Our first contribution (§6.3) introduced a data-centric approach to voice pathology detection that addresses the fundamental challenge of limited medical data availability. By combining synthetic data generation with specialized model training, we demonstrated that transformer-based models can achieve reliable pathology detection despite data constraints. The success of this approach, particularly its robustness across different languages and recording conditions, suggests promising directions for adapting foundation models to other medical domains where data scarcity is a common challenge.

The second contribution (§6.4) advanced the field through MVP, a framework for multi-source voice analysis that mirrors clinical practice. By developing specialized fusion strategies that combine information from sustained vowels and continuous speech, we achieved substantial improvements in pathology detection accuracy. The success of intermediate feature fusion, particularly with transformer-based integration, demonstrates the value of modeling cross-source relationships in medical applications. This finding has broader implications for medical AI, suggesting that explicitly modeling relationships between different diagnostic inputs can significantly improve system performance.

Our final contribution (§6.5) addressed the challenging problem of dysarthric speech recognition through a novel two-stage approach combining acoustic modeling with linguistic error correction. While achieving strong performance on continuous speech, our investigation revealed important limitations in processing isolated words. This dichotomy highlights a key challenge in medical speech technology: the need to balance general-purpose capabilities with specialized processing for specific medical conditions.

Several important themes emerge from these investigations. First, the successful application of foundation models to medical speech technology requires careful attention to domain-specific challenges. Whether through synthetic data generation, multi-source fusion, or specialized error correction, adapting these models for medical use demands more than simple fine-tuning. Second, the relationship between different types of vocal information proves crucial across all three studies. From combining different recording types in pathology detection to leveraging multiple ASR hypotheses in dysarthric speech recognition, the integration of complementary information sources consistently improves performance.

These advances in medical speech technology have significant implications for clinical practice. More accurate and robust voice analysis tools can enable earlier detection of pathologies, more precise monitoring of treatment outcomes, and better accessibility for individuals with speech impairments. However, realizing these benefits requires continued attention to the unique challenges of medical applications, from data privacy and clinical validation to the need for interpretable and reliable system behavior.

As foundation models continue to evolve, their application to medical speech technology presents both opportunities and challenges. Our work demonstrates that with careful consideration of medical requirements and constraints, these powerful models can be effectively adapted for clinical applications.

6.6.1 Future Research Directions

Our work suggests several promising directions for future research in medical speech technology.

For synthetic data generation, developing more sophisticated techniques that better capture the nuanced characteristics of pathological speech could help address data

scarcity across various medical domains. This includes exploring conditional generation methods that can produce more realistic variations of specific voice disorders and investigating ways to preserve clinically relevant features during synthesis.

In multi-source analysis, advanced fusion strategies could be extended to other types of medical data where multiple complementary sources of information are available. Future research might explore dynamic fusion mechanisms that adapt to different pathological conditions or patient characteristics. The development of interpretable fusion methods could also help clinicians better understand how different diagnostic inputs contribute to final decisions.

For dysarthric speech recognition, several critical areas demand attention. Specialized architectures for isolated word recognition could address current limitations in processing single-word utterances. More robust adaptation techniques for handling speech variability could improve performance across different severity levels and types of dysarthria. Research into personalized acoustic modeling might enable better adaptation to individual speech patterns while maintaining privacy.

Beyond these specific directions, broader challenges remain in clinical integration and validation. Future work should investigate ways to maintain model reliability across different healthcare settings, develop privacy-preserving methods for continuous model improvement, and establish protocols for clinical validation of AI-assisted diagnosis. These advances will be fundamental for translating promising research results into practical clinical tools that can meaningfully improve patient care.

Chapter 7

Conclusions

7.1 Summary of Contributions

This thesis has advanced the state-of-the-art in speech processing across four fundamental dimensions: model robustness and fairness, comprehensive evaluation frameworks, natural conversation modeling, and medical applications. Through systematic investigation and novel methodological developments, we have addressed critical gaps in current speech technology while establishing new approaches for building more reliable, trustworthy, and practical systems. Our work demonstrates that significant improvements in speech technology require advances across multiple interconnected aspects, from fundamental model architecture to practical deployment considerations.

7.1.1 Advances in Model Robustness and Fairness

Our contributions to model robustness, presented in Chapter 3, introduced five complementary approaches to analyzing and mitigating subgroup performance disparities. Each approach addresses different aspects of the fairness challenge, from discovery to mitigation, while maintaining privacy.

The divergence analysis framework represents our first major contribution in this area. Unlike previous approaches that relied on predetermined categories, our method enables systematic identification of challenging subgroups through interpretable metadata. This approach revealed that performance disparities often emerge from

complex interactions between multiple attributes, such as when gender intersects with speaking rate or when task-specific features combine with acoustic conditions. The ability of our framework to identify these subtle interaction effects provides key insights for developing more equitable systems.

Our post-processing data acquisition strategy demonstrated that targeted data selection can be more valuable than raw data quantity. By identifying and acquiring data specifically from underperforming subgroups, we achieved better results with less additional data compared to indiscriminate data collection. This finding has important implications for resource-efficient development of fair speech systems.

The in-training mitigation techniques, particularly divergence-aware regularization, proved highly effective at reducing performance disparities while maintaining overall accuracy. Our experiments showed that addressing biases during training leads to more systematic improvements than post-processing approaches. The regularization strategy's parameter efficiency and stability make it particularly suitable for practical applications.

The CLUES framework showed how contrastive learning can create more equitable internal representations. Through its three-level learning strategy, it achieves more equitable performance by reshaping the latent space of the model to better capture both task requirements and population characteristics. This approach demonstrated that addressing bias at the representation level can lead to more fundamental improvements in model fairness.

Perhaps most importantly, our privacy-preserving framework demonstrated that effective bias mitigation is possible without accessing sensitive demographic information during deployment. By leveraging confidence models to identify challenging cases, we achieved performance improvements comparable to methods that use demographic data, while maintaining user privacy. This addresses a critical barrier to deploying fair speech technology in privacy-sensitive contexts.

7.1.2 Novel Evaluation Frameworks

In Chapter 4, we developed three novel benchmarking frameworks that push speech technology evaluation beyond simple accuracy metrics. These frameworks enable more nuanced and meaningful evaluation of speech models' real-world capabilities, addressing critical gaps in current assessment methods.

SHALLOW provided the first systematic approach to quantifying and characterizing hallucinations in ASR systems. By decomposing errors into lexical, phonetic, morphological, and semantic dimensions, SHALLOW reveals nuanced failure patterns that traditional WER metrics miss entirely. Our evaluation demonstrated that even high-performing models can produce dangerous hallucinations, particularly in critical domains like healthcare transcription. The framework multi-dimensional analysis has important implications for model development, showing that architectural choices significantly influence the types of errors produced.

UnSLU-BENCH established comprehensive evaluation protocols for machine unlearning in spoken language understanding. This framework addresses the growing need for verifiable data removal from trained models, a capability increasingly required by privacy regulations. Through the introduction of the GUM metric, we provided the first holistic way to evaluate the competing objectives of forgetting effectiveness, computational efficiency, and maintained model utility. Our results demonstrated that simple approaches like Negative Gradients can be surprisingly effective when all three aspects are considered.

ARCH created a unified platform for assessing audio representation learning across diverse domains. This framework enables systematic evaluation of how well speech models generalize to other audio domains, such as music and environmental sounds. Our comprehensive evaluation revealed both the promise and limitations of current approaches, showing that pre-training data diversity often proves more important than model scale for cross-domain performance. Its modular design ensures it can evolve with the field, incorporating new datasets and models as they emerge.

7.1.3 Advances in Natural Conversational AI

Chapter 5 presented significant advances in natural conversational AI through four major contributions that address fundamental limitations in current systems. These advances create essential building blocks for more natural and inclusive conversational systems.

The voc2vec model established new state-of-the-art performance in non-verbal vocalization understanding. Pre-trained using self-supervised learning on 125 hours of carefully curated non-verbal audio data, voc2vec demonstrated significant improvements across multiple tasks, from emotion recognition to baby cry detection.

These gains stem from the ability i the model to capture subtle emotional expressions through non-verbal cues, providing a step forward to build more empathetic speech systems. voc2vec success in capturing non-verbal patterns challenges conventional approaches that focus primarily on linguistic content.

DeepDialogue provided unprecedented resources for modeling natural conversation flow. With 40,150 high-quality multi-turn dialogues spanning 41 domains and incorporating 20 distinct emotions, this dataset enables research into how emotions evolve naturally through extended interactions. Our analysis revealed several critical insights about conversation modeling: the degradation of smaller models beyond six turns, the benefits of cross-model interactions for dialogue coherence, and the importance of emotional progression in maintaining natural conversation flow. Our dual speech synthesis approach, providing both explicitly emotion-conditioned and implicitly derived emotional expressions, enables research at the intersection of text-based dialogue systems and speech-based conversational AI.

ITALIC enabled robust Italian conversational AI while preserving regional linguistic variations. As the first large-scale Italian Language Intent Classification dataset, ITALIC comprises over 16,500 utterances from 70 speakers across different regions. Our evaluation demonstrated how language-specific adaptation outperforms pure model scaling, while providing insights into the challenges of handling regional variations and acoustic conditions. The rich annotation enables research beyond intent classification, supporting tasks like speaker recognition, text-to-speech synthesis, and automatic speech recognition. Building on ITALIC's foundation, our detailed investigation into Italian language varieties demonstrated successful identification of regional speech patterns through innovative applications of contrastive learning. This work proved particularly valuable for understanding how speech technology can preserve and respect linguistic diversity rather than enforcing standardization. These advances establish important foundations for developing speech systems that can truly serve Italy's linguistically diverse population.

To complement these resources, we developed novel explainability techniques for speech classification models. Our framework combines word-level audio segment attribution with paralinguistic feature analysis, enabling detailed understanding of how models process and integrate different aspects of speech input. User studies confirmed the effectiveness of our framework in providing meaningful explana-

tions across languages and tasks, making model decisions more transparent and interpretable.

7.1.4 Medical Applications of Speech Technology

Chapter 6 demonstrated how modern speech technology can be effectively adapted for critical medical applications. Our contributions span voice pathology detection, multi-source analysis, and dysarthric speech recognition, showing both the potential and current limitations of these technologies in clinical contexts.

Our data-centric approach to voice pathology detection addressed the fundamental challenge of limited medical data availability. By combining synthetic data generation with specialized model training, we demonstrated that transformer models can achieve reliable pathology detection despite data constraints. The success of this approach, particularly its robustness across different languages and recording conditions, suggests promising directions for adapting foundation models to other medical domains where data scarcity is a common challenge.

The MVP framework advanced clinical voice analysis through sophisticated multi-source fusion techniques. By developing specialized fusion strategies that combine information from sustained vowels and continuous speech, we achieved substantial improvements in pathology detection accuracy. The success of intermediate feature fusion, particularly with transformer-based integration, demonstrates the value of modeling cross-source relationships in medical applications. This finding has broader implications for medical AI, suggesting that explicitly modeling relationships between different diagnostic inputs can significantly improve system performance.

Our work on dysarthric speech recognition introduced a novel two-stage approach combining acoustic modeling with linguistic error correction. While achieving strong performance on continuous speech, our investigation revealed important limitations in processing isolated words. This dichotomy highlights a fundamental challenge in medical speech technology: the need to balance general-purpose capabilities with specialized processing for specific medical conditions. The system's success in handling continuous speech while struggling with isolated words provides important insights for future research directions in accessibility technology.

7.2 Practical Implications

The findings and methodologies developed in this thesis have significant practical implications across multiple domains, from industry applications to healthcare and accessibility. Our work demonstrates that theoretical advances in speech technology can lead to concrete improvements in real-world applications while maintaining practical deployability.

7.2.1 Industry Applications

Our contributions provide several immediately applicable tools and frameworks for commercial speech technology development.

Bias Mitigation in Production Systems. Our privacy-preserving bias mitigation framework enables companies to improve model fairness without collecting sensitive demographic data. This approach is particularly valuable for organizations operating under strict privacy regulations like GDPR. The framework's efficiency and ease of integration make it suitable for continuous model improvement in production environments.

Quality Assurance and Testing. The SHALLOW framework provides companies with sophisticated tools for detecting and characterizing potential hallucinations before deployment. This capability is particularly relevant for high-stakes applications like medical transcription or legal documentation. UnSLU-BENCH enables systematic evaluation of privacy-preserving capabilities, an increasingly important consideration for commercial speech systems.

Enhanced User Experience. The integration of voc2vec's non-verbal understanding capabilities can significantly improve the naturalness of virtual assistants. DeepDialogue's emotional modeling framework provides templates for creating more engaging and context-aware conversational interfaces. These advances enable the development of more sophisticated and user-friendly voice-based applications.

7.2.2 Healthcare Applications

Our advances in medical speech technology have direct implications for clinical practice and healthcare delivery.

Clinical Decision Support. The MVP framework for voice pathology detection provides clinicians with sophisticated diagnostic tools that mirror clinical assessment methods. By combining multiple sources of vocal information, the system offers more comprehensive and reliable analysis than traditional single-source approaches, and enables clinicians to understand and validate system recommendations, a fundamental practice for building trust in AI-assisted diagnosis.

Patient Monitoring. Our approaches to voice analysis enable more frequent and consistent monitoring of voice disorders. The ability to detect subtle changes in voice quality supports earlier intervention and more precise treatment adjustment. The non-invasive and automated nature of our methods makes them particularly suitable for remote monitoring applications.

Clinical Documentation. Our advances in dysarthric speech recognition directly support more accurate and efficient clinical documentation. The system's ability to handle continuous speech with high accuracy makes it practical for real-world clinical use. The robustness of the two-stage approach to different speaking patterns helps ensure reliable performance across diverse patient populations.

Research Tools. The release of standardized datasets and evaluation frameworks supports clinical research in voice disorders. Our multi-source fusion techniques provide new methodologies for studying the relationship between different vocal tasks. These tools enable more rigorous and reproducible research in clinical speech science.

7.2.3 Accessibility and Inclusion

Our work makes significant contributions to technology accessibility and inclusion.

Language Diversity. The ITALIC dataset and associated research advances support better speech technology for Italian speakers. Our work on regional variations helps preserve linguistic diversity rather than enforcing standardization. These advances provide templates for developing inclusive speech technology in other languages.

Speech Disorders. Our dysarthric speech recognition system significantly improves technology access for individuals with motor speech disorders. The strong performance on continuous speech enables more natural interaction with voice-controlled

interfaces. Our findings about isolated word recognition challenges guide future research priorities in accessibility technology.

Emotional Intelligence. The integration of non-verbal understanding and emotional modeling enables more natural interaction for diverse user groups. The emotional framework in DeepDialogue supports the development of more empathetic conversational systems. These capabilities are particularly valuable for applications serving users with different communication styles and needs.

Privacy Protection. Our privacy-preserving approaches ensure that accessibility improvements don't come at the cost of user privacy. The ability to improve model fairness without collecting sensitive data supports broader technology adoption. These advances help create more inclusive technology while protecting user rights.

7.2.4 Research Community Impact

Our contributions provide valuable resources and methodologies for the broader research community.

Open Resources. The release of multiple datasets (DeepDialogue, ITALIC) enables new research directions in conversational AI and multilingual speech processing. Our novel foundation models make significant contributions to the field: voc2vec provides the first specialized model for non-verbal vocalization understanding, while our AudioSet-pretrained SSL models (wav2vec 2.0, HuBERT) extend the capabilities of speech models to general audio processing. Our evaluation frameworks (SHALLOW, UnSLU-BENCH, ARCH) provide standardized tools for assessing critical model capabilities. The publication of all models, implementation details, and code supports reproducibility and further development. These open-source contributions have already enabled numerous follow-up studies and practical applications, demonstrating their value to the research community.

Methodological Advances. Our systematic approach to analyzing model behavior provides templates for investigating other aspects of speech technology. The demonstration of effective privacy-preserving techniques establishes new standards for responsible AI development. Our fusion strategies for combining different information sources offer insights for other multimodal learning problems.

Benchmark Standards. ARCH establishes new standards for evaluating audio representation learning across diverse domains. SHALLOW provides the first comprehensive framework for analyzing hallucinations in speech recognition. UnSLU-BENCH introduces rigorous evaluation protocols for machine unlearning in speech technology.

Interdisciplinary Integration. Our work demonstrates effective integration of insights from linguistics, clinical practice, and machine learning. The success of this interdisciplinary approach provides models for future research collaborations. Our findings highlight the value of combining domain expertise with technical innovation.

7.3 Limitations and Future Directions

While our work has made significant advances, several important limitations and opportunities for future research remain. Understanding these limitations helps identify relevant directions for continued development of robust and responsible speech technology.

7.3.1 Model Robustness and Fairness

Despite our advances in bias mitigation, several challenges remain.

Generalization Across Domains. Our current approaches show strong performance within specific datasets, but their effectiveness across extremely different domains requires further investigation. Future work should explore more sophisticated transfer learning techniques to adapt fairness improvements across diverse application areas.

Dynamic Bias Detection. As models are deployed in real-world settings, new biases may emerge that were not present in the original training data. Research is needed on automated methods for identifying and characterizing these emerging biases during deployment. This could involve developing online learning techniques that continuously monitor and adapt to changing performance patterns across subgroups.

Causal Understanding. While our methods effectively identify and mitigate performance disparities, a deeper causal understanding of why these disparities occur remains elusive. Future work could investigate causal inference techniques to dis-

entangle the complex factors contributing to model bias. This could in turn lead to more targeted and effective mitigation strategies.

Multi-Objective Optimization. Our current approaches primarily focus on reducing performance disparities while maintaining overall accuracy. Future research should explore more sophisticated multi-objective optimization techniques that simultaneously consider fairness, accuracy, computational efficiency, and other important criteria.

7.3.2 Evaluation Frameworks

Our evaluation frameworks provide valuable tools for assessing speech models, but several areas for improvement and expansion remain.

Multilingual Hallucination Analysis. SHALLOW currently focuses on English ASR systems. Extending the framework to multilingual settings could reveal important insights about how hallucination patterns vary across languages and cultural contexts. This expansion would require careful consideration of language-specific linguistic features and cultural norms.

Efficient Unlearning at Scale. As language models continue to grow in size, the computational cost of unlearning becomes increasingly prohibitive. Research is needed on more efficient unlearning techniques that can scale to models with billions or even trillions of parameters. This could involve developing approximate unlearning methods that provide strong privacy guarantees with minimal performance impact.

Task-Specific Audio Benchmarks. While ARCH provides a comprehensive evaluation of general audio representations, there is a need for more specialized benchmarks targeting specific application areas. Future work could develop domain-specific evaluation suites for areas like medical diagnosis or environmental monitoring.

Human-AI Collaborative Evaluation. Our current frameworks focus primarily on automated evaluation metrics. Developing evaluation methodologies that incorporate human judgment alongside computational metrics could provide more holistic assessments of model performance. This is particularly important for subjective tasks like assessing the naturalness of synthesized speech or the appropriateness of emotional responses in dialogue systems.

7.3.3 Natural Conversation Modeling

While our contributions significantly advance conversational AI, several challenges remain.

Long-Term Coherence. Current models struggle to maintain coherence and context over extended multi-turn conversations. Research is needed on architectural innovations and training techniques that can capture and utilize long-range dependencies in dialogue.

Multimodal Integration. Real-world conversations involve more than just speech and text. Future work should explore how to effectively integrate non-verbal vocalizations with other communication channels like facial expressions, gestures, and body language.

Cross-Cultural Adaptation. Developing truly global conversational AI systems requires more than just multilingual capabilities. Research is needed on efficiently adapting models across cultures while preserving local conversational norms and communication styles.

Ethical Considerations. As conversational AI becomes more sophisticated, ethical concerns around deception and human attachment to AI systems become increasingly relevant. Future work must address these ethical challenges alongside technical developments.

7.3.4 Medical Applications

Our work in medical speech technology reveals several important areas for future research.

Explainable AI for Clinical Use. While our current approaches provide some level of interpretability, further work is needed to develop truly explainable AI systems suitable for high-stakes medical decisions. This could involve developing novel visualization techniques or natural language explanations tailored for clinical users.

Multimodal Medical Analysis. Our multi-source fusion approach for voice pathology detection could be extended to incorporate other types of medical data. Future research could explore how to effectively combine speech analysis with electronic

health records, imaging data, or physiological sensors for more comprehensive diagnostic tools.

Longitudinal Studies. To fully validate the clinical utility of our approaches, long-term studies tracking patient outcomes are necessary. This would involve collaborations with healthcare providers to deploy and monitor these technologies in real clinical settings over extended periods.

Personalized Speech Technology. Given the high variability in speech patterns among individuals with disorders like dysarthria, research into rapidly adaptable or personalizable models could significantly improve performance. This might involve developing few-shot learning techniques specifically tailored for clinical speech applications.

7.4 Concluding Thoughts

This thesis has demonstrated that building truly robust, responsible, and trustworthy speech technology requires advances across multiple interconnected dimensions. From improving model fairness to developing comprehensive evaluation frameworks, from enabling more natural conversation to advancing medical applications, each contribution addresses fundamental challenges in modern speech processing.

The success of our approaches shows that careful consideration of real-world constraints and requirements can lead to significant improvements in speech technology. Our work establishes new methodologies for developing more inclusive and reliable systems while maintaining practical deployability. The integration of insights from linguistics, clinical practice, and machine learning demonstrates the power of interdisciplinary approaches in tackling complex technological challenges.

As speech technology continues to evolve and impact more aspects of daily life, the principles and approaches developed in this thesis provide valuable foundations for future advances. The combination of technical innovation with responsible development practices demonstrated here offers a template for creating speech technology that better serves all users while maintaining high standards of privacy and fairness.

Looking ahead, several key themes emerge as central to the future development of speech technology.

Holistic evaluation. As models become more sophisticated, evaluation must go beyond simple accuracy metrics to consider fairness, robustness, privacy, and real-world applicability. Our frameworks provide starting points, but continuous refinement of evaluation methodologies will be essential.

Ethical AI development. The potential impact of speech technology on privacy, accessibility, and human interaction demands careful ethical consideration at every stage of development. Future work must prioritize responsible innovation that respects user rights and societal values.

Interdisciplinary collaboration. The most significant advances often arise at the intersection of different fields. Continued collaboration between speech technologists, linguists, clinicians, and ethicists will be crucial for addressing complex challenges.

Adaptable and personalized systems. As speech technology enters more diverse application areas, the ability to rapidly adapt to new domains and individual user needs becomes increasingly important. Developing flexible, personalized systems while maintaining privacy and fairness presents exciting research opportunities.

The future of speech technology lies not just in improving accuracy metrics, but in creating systems that are truly robust, responsible, and trustworthy. This thesis has established important steps toward that goal while highlighting fundamental directions for continued research and development. By building on these foundations and embracing the challenges ahead, we can work towards a future where speech technology enhances human communication and capabilities in ways that are both powerful and ethically sound.

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[2] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

[3] Pranav Dheram, Murugesan Ramakrishnan, Anirudh Raju, I-Fan Chen, Brian King, Katherine Powell, Melissa Saboowala, Karan Shetty, and Andreas Stolcke. Toward fairness in speech recognition: Discovery and mitigation of performance disparities. In *Proc. Interspeech 2022*, pages 1268–1272, 2022.

[4] Irina-Elena Veliche and Pascale Fung. Improving fairness and robustness in end-to-end speech recognition through unsupervised clustering. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[5] Moreno La Quatra, Juan Rafael Orozco-Arroyave, and Marco Sabato Siniscalchi. Bilingual dual-head deep model for parkinson's disease detection from speech. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025.

[6] Dayana Ribas, Miguel A Pastor, Antonio Miguel, David Martínez, Alfonso Ortega, and Eduardo Lleida. Automatic voice disorder detection using self-supervised representations. *IEEE Access*, 2023.

[7] Rita Frieske and Bertram E Shi. Hallucinations in neural automatic speech recognition: Identifying errors and hallucinatory models. *arXiv preprint arXiv:2401.01572*, 2024.

[8] Chunxi Liu, Michael Picheny, Leda Sarı, Pooja Chitkara, Alex Xiao, Xiaohui Zhang, Mark Chou, Andres Alvarado, Caner Hazirbas, and Yatharth Saraf. Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[9] Ayush K Tarun, Vikram S Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[10] Emmie Hine, Claudio Novelli, Mariarosaria Taddeo, and Luciano Floridi. Supporting trustworthy ai through machine unlearning. *Science and Engineering Ethics*, 30(5):43, 2024.

[11] Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X Mei, Hilke Schellmann, and Mona Sloane. Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1672–1681, 2024.

[12] Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. Machine unlearning: A survey. *ACM Comput. Surv.*, 56(1), August 2023.

[13] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[14] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *Interspeech*, 2022.

[15] Eliana Pastor, Luca de Alfaro, and Elena Baralis. Looking for trouble: Analyzing classifier behavior via pattern divergence. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, page 1400–1412. ACM, 2021.

[16] Alkis Koudounas, Eliana Pastor, Giuseppe Attanasio, Vittorio Mazzia, Manuel Giollo, Thomas Gueudre, Luca Cagliero, Luca de Alfaro, Elena Baralis, and Daniele Amberti. Exploring subgroup performance in end-to-end speech models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[17] Alkis Koudounas, Eliana Pastor, Giuseppe Attanasio, Vittorio Mazzia, Manuel Giollo, Thomas Gueudre, Elisa Reale, Luca Cagliero, Sandro Cumani, Luca de Alfaro, Elena Baralis, and Daniele Amberti. Towards comprehensive subgroup performance analysis in speech models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1468–1480, 2024.

[18] Alkis Koudounas, Eliana Pastor, Elena Baralis, et al. Assessing speech model performance: A subgroup perspective. In *SEBD 2024: 32nd Symposium on Advanced Database System*, volume 3741, pages 101–111. CEUR Workshop Proceedings, 2024.

[19] Alkis Koudounas, Eliana Pastor, and Elena Baralis. Assessing and mitigating speech model biases via pattern mining. In *KDD PhD COnsortium*, 2024.

[20] Alkis Koudounas, Eliana Pastor, Luca de Alfaro, and Elena Baralis. Mitigating subgroup disparities in speech models: A divergence-aware dual strategy. *IEEE Transactions on Audio, Speech and Language Processing*, 33:883–895, 2025.

[21] Alkis Koudounas, Eliana Pastor, Giuseppe Attanasio, Luca, Luca de Alfaro, and Elena Baralis. Prioritizing data acquisition for end-to-end speech model improvement. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2024.

[22] Alkis Koudounas, Flavio Giobergia, Eliana Pastor, and Elena Baralis. A contrastive learning approach to mitigate bias in speech models. In *Proc. Interspeech 2024*, pages 827–831, 2024.

[23] Alkis Koudounas, Eliana Pastor, Vittorio Mazzia, Manuel Giollo, Thomas Gueudre, Elisa Reale, Giuseppe Attanasio, Luca Cagliero, Sandro Cumani, Luca De Alfaro, Elena Baralis, and Daniele Amberti. Leveraging confidence models for identifying challenging data subgroups in speech models. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 134–138, 2024.

[24] Alkis Koudounas, Eliana Pastor, Vittorio Mazzia, Manuel Giollo, Thomas Gueudre, Elisa Reale, Luca Cagliero, Sandro Cumani, Luca De Alfaro, Elena Baralis, and Daniele Amberti. Privacy preserving data selection for bias mitigation in speech models. In Georg Rehm and Yunyao Li, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 738–748, Vienna, Austria, July 2025. Association for Computational Linguistics.

[25] Alkis Koudounas, Moreno La Quatra, Manuel Giollo, Sabato Marco Siniscalchi, and Elena Baralis. Hallucination benchmark for speech foundation models. *arXiv preprint arXiv:2510.16567*, 2025.

[26] Alkis Koudounas, Claudio Savelli, Flavio Giobergia, and Elena Baralis. "Alexa, can you forget me?" Machine Unlearning Benchmark in Spoken Language Understanding. In *Interspeech 2025*, pages 1768–1772, 2025.

[27] Moreno La Quatra, Alkis Koudounas, Lorenzo Vaiani, Elena Baralis, Paolo Garza, Luca Cagliero, and Sabato Marco Siniscalchi. Benchmarking representations for speech, music, and acoustic events. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024.

[28] Alkis Koudounas, Moreno La Quatra, Sabato Marco Siniscalchi, and Elena Baralis. voc2vec: A foundation model for non-verbal vocalization. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025.

[29] Alkis Koudounas, Moreno La Quatra, and Elena Baralis. Deepdialogue: A multi-turn emotionally-rich spoken dialogue dataset. *arXiv preprint arXiv:2505.19978*, 2025.

[30] Alkis Koudounas, Moreno La Quatra, Lorenzo Vaiani, Luca Colomba, Giuseppe Attanasio, Eliana Pastor, Luca Cagliero, and Elena Baralis. ITALIC: An Italian Intent Classification Dataset. In *Proc. INTERSPEECH 2023*, pages 2153–2157, 2023.

[31] Moreno La Quatra, Alkis Koudounas, Elena Baralis, and Sabato Marco Siniscalchi. Speech analysis of language varieties in Italy. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15147–15159, 2024.

[32] Eliana Pastor, Alkis Koudounas, Giuseppe Attanasio, Dirk Hovy, and Elena Baralis. Explaining speech classification models via word-level audio segments and paralinguistic features. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2221–2238, 2024.

[33] Alkis Koudounas, Gabriele Ciravegna, Marco Fantini, Erika Crosetti, Giovanni Succo, Tania Cerquitelli, and Elena Baralis. Voice disorder analysis: a transformer-based approach. *Interspeech 2024*, 2024.

[34] Alkis Koudounas, Moreno La Quatra, Gabriele Ciravegna, Marco Fantini, Erika Crosetti, Giovanni Succo, Tania Cerquitelli, Sabato Marco Siniscalchi, and Elena Baralis. MVP: Multi-source Voice Pathology detection. In *Interspeech 2025*, pages 3548–3552, 2025.

[35] Moreno La Quatra, Alkis Koudounas, Valerio Mario Salerno, and Sabato Marco Siniscalchi. Exploring Generative Error Correction for Dysarthric Speech Recognition. In *Interspeech 2025*, pages 3284–3288, 2025.

[36] Irene Benedetto, Alkis Koudounas, Lorenzo Vaiani, Eliana Pastor, Elena Baralis, Luca Cagliero, and Francesco Tarasconi. PoliToHFI at SemEval-2023 task 6: Leveraging entity-aware and hierarchical transformers for legal entity recognition and court judgment prediction. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1401–1411, Toronto, Canada, July 2023. Association for Computational Linguistics.

[37] Irene Benedetto, Alkis Koudounas, Lorenzo Vaiani, Eliana Pastor, Luca Cagliero, Francesco Tarasconi, and Elena Baralis. Boosting court judgment prediction and explanation using legal entities. *Artificial Intelligence and Law*, pages 1–36, 2024.

[38] Irene Benedetto, Alkis Koudounas, Lorenzo Vaiani, Eliana Pastor, Luca Cagliero, and Francesco Tarasconi. MAINDZ at SemEval-2024 task 5: CLUEDO - choosing legal oUtcome by explaining decision through oversight. In Atul Kr. Ojha, A. Seza Doğruöz, Harish Tayyar Madabushi, Giovanni Da San Martino, Sara Rosenthal, and Aiala Rosá, editors, *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 997–1005, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[39] Alkis Koudounas, Flavio Giobergia, Irene Benedetto, Simone Monaco, Luca Cagliero, Daniele Apiletti, Elena Baralis, et al. baρtti at geolingit: Beyond boundaries, enhancing geolocation prediction and dialect classification on social media in italy. In *CEUR workshop proceedings*. CEUR, 2023.

[40] Flavio Giobergia, Alkis Koudounas, and Elena Baralis. Large language models-aided literature reviews: A study on few-shot relevance classification. In *2024 IEEE 18th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–5, 2024.

[41] Luca Cagliero, Lorenzo Vaiani, Eliana Pastor, Alkis Koudounas, Elena Baralis, Vittorio Mazzia, Sandro Pollastrini, Thomas Gueudre, Manuel Giollo, Daniele Amberti, and Yue Wu. Detecting and mitigating challenges in zero-shot video summarization with video LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 286–301, Vienna, Austria, July 2025. Association for Computational Linguistics.

[42] Federico Borra, Claudio Savelli, Giacomo Rosso, Alkis Koudounas, and Flavio Giobergia. Malto at semeval-2024 task 6: Leveraging synthetic data for llm hallucination detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1678–1684, 2024.

[43] Claudio Savelli, Alkis Koudounas, and Flavio Giobergia. MALTO at SemEval-2025 task 3: Detecting hallucinations in LLMs via uncertainty quantification and larger model validation. In Sara Rosenthal, Aiala Rosá, Debanjan Ghosh, and Marcos Zampieri, editors, *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1318–1324, Vienna, Austria, July 2025. Association for Computational Linguistics.

[44] Giuseppe Concialdi, Alkis Koudounas, Eliana Pastor, Barbara Di Eugenio, and Elena Baralis. Ainur: Harmonizing speed and quality in deep music generation through lyrics-audio embeddings. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1146–1150. IEEE, 2024.

[45] Moreno La Quatra, Lorenzo Vaiani, Alkis Koudounas, Luca Cagliero, Paolo Garza, and Elena Baralis. How much attention should we pay to mosquitoes? In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7135–7139, 2022.

[46] Flavio Giobergia, Alkis Koudounas, and Elena Baralis. Reconstructing atmospheric parameters of exoplanets using deep learning. In *2023 IEEE 17th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–6, 2023.

[47] Alkis Koudounas, Flavio Giobergia, and Elena Baralis. Bad exoplanet! explaining degraded performance when reconstructing exoplanets atmospheric parameters. In *NeurIPS 2023 AI for Science Workshop*, 2023.

[48] Alkis Koudounas, Flavio Giobergia, and Elena Baralis. Ex (o) plain: Subgroup-level analysis of exoplanet atmospheric parameters. *IEEE Access*, 2024.

[49] Alkis Koudounas and Flavio Giobergia. Houston we have a divergence: A subgroup performance analysis of asr models. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 812–813. IEEE, 2024.

[50] Alkis Koudounas, Flavio Giobergia, and Elena Baralis. Time-of-flight cameras in space: Pose estimation with deep learning methodologies. In *2022 IEEE 16th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–6, 2022.

[51] Lia Morra, Alberto Azzari, Letizia Bergamasco, Marco Braga, Luigi Capogrosso, Federico Delrio, Giuseppe Di Giacomo, Simone Eiraudo, Giorgia Ghione, Rocco Giudice, et al. Designing logic tensor networks for visual sudoku puzzle classification. In *NeSy*, pages 223–232, 2023.

[52] Flavio Giobergia, Claudio Savelli, Alkis Koudounas, Elena Baralis, et al. Quantum feature selection from interpretable models using qubo formulation. *Working Notes of CLEF*, 2025.

[53] Alkis Koudounas and Simone Fiori. Gradient-based learning methods extended to smooth manifolds applied to automated clustering. *Journal of Artificial Intelligence Research*, 68:777–816, 2020.

[54] Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, et al. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210, 2022.

[55] Takanori Ashihara, Takafumi Moriya, Kohei Matsuura, Tomohiro Tanaka, Yusuke Ijima, Taichi Asami, Marc Delcroix, and Yukinori Honma. Speechglue: How well can self-supervised speech models capture linguistic knowledge? In *Proc. Interspeech 2023*, pages 2888–2892, 2023.

[56] Takanori Ashihara, Marc Delcroix, Takafumi Moriya, Kohei Matsuura, Taichi Asami, and Yusuke Ijima. What do self-supervised speech and speaker models learn? new findings from a cross model layer-wise analysis. In *ICASSP*

2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 10166–10170. IEEE, 2024.

[57] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[58] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[59] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[60] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France, May 2020. European Language Resources Association.

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[62] Mark JF Gales, Kate M Knill, Anton Ragni, and Shakti P Rath. Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. In *Fourth International workshop on spoken language technologies for under-resourced languages (SLTU-2014)*, pages 16–23. International Speech Communication Association (ISCA), 2014.

[63] Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online, August 2021. Association for Computational Linguistics.

[64] Mateusz Barański, Jan Jasiński, Julitta Bartolewska, Stanisław Kacprzak, Marcin Witkowski, and Konrad Kowalczyk. Investigation of whisper asr hallucinations induced by non-speech audio. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[65] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proc. of the National Academy of Sciences*, 117(14):7684–7689, 2020.

[66] K Knill and S Young. Hidden markov models in speech and language processing. In *Corpus-based methods in language and speech processing*, pages 27–68. Springer, 1997.

[67] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Speech parameter generation algorithms for hmm-based speech synthesis. In *2000 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 00CH37100)*, volume 3, pages 1315–1318. IEEE, 2000.

[68] Dong Yu and Lin Deng. *Automatic speech recognition*, volume 1. Springer, 2016.

[69] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. Deep speech: Scaling up end-to-end speech recognition, 2014.

[70] Jinyu Li et al. Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, 11(1), 2022.

[71] Rohit Prabhavalkar, Takaaki Hori, Tara N Sainath, Ralf Schlüter, and Shinji Watanabe. End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:325–351, 2023.

[72] Hainan Xu, Fei Jia, Somshubra Majumdar, He Huang, Shinji Watanabe, and Boris Ginsburg. Efficient sequence transduction by jointly predicting tokens and durations. In *International Conference on Machine Learning*, pages 38462–38484. PMLR, 2023.

[73] Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, et al. Fast conformer with linearly scalable attention for efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023.

[74] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. In *Proc. Interspeech 2020*, pages 5036–5040, 2020.

[75] Suyoun Kim, Abhinav Arora, Duc Le, Ching-Feng Yeh, Christian Fuegen, Ozlem Kalinli, and Michael L. Seltzer. Semantic distance: A new metric for asr performance analysis towards spoken language understanding. In *Interspeech 2021*, pages 1977–1981, 2021.

[76] Bornali Phukon, Xiuwen Zheng, and Mark Hasegawa-Johnson. Aligning ASR Evaluation with Human and LLM Judgments: Intelligibility Metrics Using Phonetic, Semantic, and NLI Approaches. In *Interspeech*, 2025.

[77] Gokhan Tur and Renato De Mori. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons, 2011.

[78] Dmitriy Serdyuk, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua Bengio. Towards end-to-end spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5754–5758. IEEE, 2018.

[79] Alkis Koudounas, Moreno La Quatra, Eliana Pastor, Sabato Marco Siniscalchi, and Elena Baralis. "KAN you hear me?" Exploring Kolmogorov-Arnold Networks for Spoken Language Understanding. In *Interspeech 2025*, pages 4123–4127, 2025.

[80] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[81] Eric Goldman. An introduction to the california consumer privacy act (ccpa). *Santa Clara Univ. Legal Studies Research Paper*, 2020.

[82] Je Hun Jeon, Rui Xia, and Yang Liu. Sentence level emotion recognition based on decisions from subsentence segments. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4940–4943, 2011.

[83] Wenjing Han, Huabin Ruan, Xiaomin Chen, Zhixiang Wang, Haifeng Li, and Björn Schuller. Towards temporal modelling of categorical speech emotion recognition. In *Interspeech 2018*, pages 932–936, 2018.

[84] Xiaomin Chen, Wenjing Han, Huabin Ruan, Jiamu Liu, Haifeng Li, and Dongmei Jiang. Sequence-to-sequence modelling for categorical speech emotion recognition using recurrent neural network. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pages 1–6, 2018.

[85] Yue Xie, Ruiyu Liang, Zhenlin Liang, Chengwei Huang, Cairong Zou, and Björn Schuller. Speech emotion classification using attention-based lstm. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(11):1675–1685, 2019.

[86] James A Russell. Affective space is bipolar. *Journal of personality and social psychology*, 37(3):345, 1979.

[87] Eddie Harmon-Jones, Cindy Harmon-Jones, and Elizabeth Summerell. On the importance of both dimensional and discrete models of emotion. *Behavioral sciences*, 7(4):66, 2017.

[88] Aneesha Sampath, James Tavernor, and Emily Mower Provost. Efficient fine-tuning for dimensional speech emotion recognition in the age of transformers. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[89] Disa A Sauter, Frank Eisner, Andrew J Calder, and Sophie K Scott. Perceptual cues in nonverbal vocal expressions of emotion. *Quarterly journal of experimental psychology*, 63(11):2251–2272, 2010.

[90] César F Lima, São Luís Castro, and Sophie K Scott. When voices get emotional: A corpus of nonverbal vocalizations for research on emotion processing. *Behavior research methods*, 45(4):1234–1245, 2013.

[91] Disa A Sauter, Frank Eisner, Paul Ekman, and Sophie K Scott. Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, 107(6):2408–2412, 2010.

[92] Sven Grawunder, Natalie Uomini, Liran Samuni, Tatiana Bortolato, Cédric Girard-Buttoz, Roman M Wittig, and Catherine Crockford. Chimpanzee vowel-like sounds and voice quality suggest formant space expansion through the hominoid lineage. *Philosophical Transactions of the Royal Society B*, 377(1841):20200455, 2022.

[93] Gabriel Jorgewich-Cohen, Simon William Townsend, Linilson Rodrigues Padovese, Nicole Klein, Peter Praschag, Camila R Ferrara, Stephan Ettmar, Sabrina Menezes, Arthur Pinatti Varani, Jaren Serano, et al. Common evolutionary origin of acoustic communication in choanate vertebrates. *Nature Communications*, 13(1):6089, 2022.

[94] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.

[95] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

[96] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

[97] Carlos Busso, Reza Lotfian, Kusha Sridhar, Ali N Salman, Wei-Cheng Lin, Lucas Goncalves, Srinivas Parthasarathy, Abinay Reddy Naini, Seong-Gyun Leem, Luz Martinez-Lucas, et al. The msp-podcast corpus. *arXiv preprint arXiv:2509.09791*, 2025.

[98] Jing Peng, Yucheng Wang, Bohan Li, Yiwei Guo, Hankun Wang, Yangui Fang, Yu Xi, Haoyu Li, Xu Li, Ke Zhang, et al. A survey on speech large language models for understanding. *Authorea Preprints*, 2025.

[99] Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. On the landscape of spoken language models: A comprehensive survey. *arXiv preprint arXiv:2504.08528*, 2025.

[100] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, MA Zejun, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[101] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

[102] Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

[103] Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, et al. Textually pretrained speech language models. *Advances in Neural Information Processing Systems*, 36:63483–63501, 2023.

[104] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.

[105] Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738, 2022.

[106] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

[107] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

[108] Yuanyuan Zhang, Yixuan Zhang, Bence Mark Halpern, Tanvina Patel, and Odette Scharenborg. Mitigating bias against non-native accents. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2022, pages 3168–3172, 2022.

[109] Rachael Tatman and Conner Kasten. Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions. In *Interspeech*, pages 934–938, 2017.

[110] Joshua L. Martin and Kevin Tang. Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual "be". In *Proc. Interspeech 2020*, pages 626–630, 2020.

[111] Leda Sarı, Mark Hasegawa-Johnson, and Chang D. Yoo. Counterfactually fair automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3515–3525, 2021.

[112] Mahault Garnerin, Solange Rossato, and Laurent Besacier. Investigating the impact of gender representation in asr training data: A case study on librispeech. In *3rd Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 2021.

[113] Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*, 2021.

[114] Zhe Liu, Irina-Elena Veliche, and Fuchun Peng. Model-based approach for measuring the fairness in asr. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6532–6536. IEEE, 2022.

[115] Joan Palmiter Bajorek. Voice recognition still has significant race and gender biases. *Harvard Business Review*, 10, 2019.

[116] Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. "i don't think these devices are very culturally sensitive."—impact of automated speech recognition errors on african americans. *Frontiers in Artificial Intelligence*, page 169, 2021.

[117] Eliana Pastor, Andrew Gavgavian, Elena Baralis, and Luca de Alfaro. How divergent is your data? *Proc. VLDB Endow.*, 14(12):2835–2838, jul 2021.

[118] Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Euijong Whang. Automated data slicing for model validation: A big data - ai integration approach. *IEEE TKDE*, 32(12):2284–2296, 2020.

[119] Svetlana Sagadeeva and Matthias Boehm. SliceLine: Fast, linear-algebra-based slice finding for ML model debugging. In *SIGMOD/PODS '21*, page 2290–2299, 2021.

[120] Oliver Niebuhr and Alexis Michaud. Speech data acquisition: the underestimated challenge. *KALIPHO-Kieler Arbeiten zur Linguistik und Phonetik*, 3:1–42, 2015.

[121] Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *Advances in neural information processing systems*, 31, 2018.

[122] Ki Hyun Tae and Steven Euijong Whang. Slice tuner: A selective data acquisition framework for accurate and fair machine learning models. In *Proceedings of the 2021 International Conference on Management of Data*, pages 1771–1783, 2021.

[123] Abolfazl Asudeh, Zhongjun Jin, and HV Jagadish. Assessing and remedying coverage for a given dataset. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 554–565. IEEE, 2019.

[124] Yuanyuan Zhang, Aaricia Herygers, Tanvina Patel, Zhengjun Yue, and Odette Scharenborg. Exploring data augmentation in bias mitigation against non-native-accented speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023.

[125] Jianfeng Chi, William Shand, Yaodong Yu, Kai-Wei Chang, Han Zhao, and Yuan Tian. Conditional supervised contrastive learning for fair text classification. In *EMNLP*, pages 2736–2756, 2022.

[126] Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. Contrastive learning for fair representations. *arXiv preprint arXiv:2109.10645*, 2021.

[127] Youngkyu Hong and Eunho Yang. Unbiased classification through bias-contrastive and bias-balanced learning. *NeurIPS*, 2021.

[128] Minwoo Lee, Hyukhun Koh, Kang-il Lee, Dongdong Zhang, Minsung Kim, and Kyomin Jung. Target-agnostic gender-aware contrastive learning for mitigating bias in multilingual machine translation. *arXiv preprint arXiv:2305.14016*, 2023.

[129] Haider Al-Tahan and Yalda Mohsenzadeh. Clar: Contrastive learning of auditory representations. In *International Conference on Artificial Intelligence and Statistics*, pages 2530–2538. PMLR, 2021.

[130] Rong Ye, Mingxuan Wang, and Lei Li. Cross-modal contrastive learning for speech translation. In *NAACL HLT*, 2022.

[131] Nik Vaessen and David A van Leeuwen. The effect of batch size on contrastive self-supervised speech representation learning. *arXiv preprint arXiv:2402.13723*, 2024.

[132] Tao Han, Hantao Huang, Ziang Yang, and Wei Han. Supervised contrastive learning for accented speech recognition. *arXiv preprint arXiv:2107.00921*, 2021.

[133] Minh Tran and Mohammad Soleymani. Privacy-preserving representation learning for speech understanding. In *Proc. Interspeech*, 2023.

[134] Shihao Chen, Liping Chen, Jie Zhang, KongAik Lee, Zhenhua Ling, and Lirong Dai. Adversarial speech for voice privacy protection from personalized speech generation. In *ICASSP*, 2024.

[135] Kei Hashimoto, Junichi Yamagishi, and Isao Echizen. Privacy-preserving sound to degrade automatic speaker verification performance. In *ICASSP*, 2016.

[136] Michele Panariello, Francesco Nespoli, Massimiliano Todisco, and Nicholas Evans. Speaker anonymization using neural audio codec language models. In *ICASSP*, 2024.

[137] Hua Shen, Yuguang Yang, Guoli Sun, Ryan Langman, Eunjung Han, Jasha Droppo, and Andreas Stolcke. Improving fairness in speaker verification via group-adapted fusion network. In *ICASSP*. IEEE, 2022.

[138] Woojay Jeon, Maxwell Jordan, and Mahesh Krishnamoorthy. On modeling asr word confidence. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6324–6328. IEEE, 2020.

[139] Amber Afshan, Kshitiz Kumar, and Jian Wu. Sequence-level confidence classifier for asr utterance accuracy and application to acoustic models. *Interspeech*, 2021.

[140] David Qiu, Qiujia Li, Yanzhang He, Yu Zhang, Bo Li, Liangliang Cao, Rohit Prabhavalkar, Deepti Bhatia, Wei Li, Ke Hu, et al. Learning word-level confidence for subword end-to-end asr. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6393–6397. IEEE, 2021.

[141] Qiujia Li, David Qiu, Yu Zhang, Bo Li, Yanzhang He, Philip C Woodland, Liangliang Cao, and Trevor Strohman. Confidence estimation for attention-based sequence-to-sequence models for speech recognition. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6388–6392. IEEE, 2021.

[142] Qiujia Li, Yu Zhang, David Qiu, Yanzhang He, Liangliang Cao, and Philip C Woodland. Improving confidence estimation on out-of-domain data for end-to-end speech recognition. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6537–6541. IEEE, 2022.

[143] Jan Niehues and Ngoc-Quan Pham. Modeling confidence in sequence-to-sequence models. *arXiv preprint arXiv:1910.01859*, 2019.

[144] Andrew F. Siegel. Chapter 10 - hypothesis testing: Deciding between reality and coincidence. In Andrew F. Siegel, editor, *Practical Business Statistics (Sixth Edition)*, pages 249–287. Springer Science & Business Media, sixth edition edition, 2012.

[145] Lloyd S Shapley et al. A value for n-person games. *Contributions to the Theory of Games*, 1953.

[146] Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. Speech model pre-training for end-to-end spoken language understanding. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*, pages 814–818, 2019.

[147] Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. SLURP: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7252–7262, 2020.

[148] Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198, 2021.

[149] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *EMNLP: System Demonstrations*, October 2020.

[150] J.F. Hair, W.C. Black, B.J. Babin, and R.E. Anderson. *Multivariate Data Analysis*. Cengage, 2019.

[151] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online, July 2020. Association for Computational Linguistics.

[152] Stephen M Chu and Daniel Povey. Speaking rate adaptation using continuous frame rate normalization. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010.

[153] Craig S Greenberg, Désiré Bansé, George R Doddington, Daniel Garcia-Romero, John J Godfrey, Tomi Kinnunen, Alvin F Martin, Alan McCree, Mark Przybocki, and Douglas A Reynolds. The nist 2014 speaker recognition

i-vector machine learning challenge. In *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[154] Fuchuan Tong, Siqi Zheng, Haodong Zhou, Xingjia Xie, Qingyang Hong, and Lin Li. Deep representation decomposition for rate-invariant speaker verification. *arXiv preprint arXiv:2205.14294*, 2022.

[155] Rishikesh Magar and Amir Barati Farimani. Learning from mistakes: Sampling strategies to efficiently train machine learning models for material property prediction. *Computational Materials Science*, 224, 2023.

[156] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[157] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *CVPR*, 2019.

[158] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[159] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[160] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 2021.

[161] Prakhar Swarup, Roland Maas, Sri Garimella, Sri Harish Mallidi, and Björn Hoffmeister. Improving ASR Confidence Scores for Alexa Using Acoustic and Hypothesis Embeddings. In *Proc. Interspeech*, 2019.

[162] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

[163] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

[164] Marianna Martindale, Marine Carpuat, Kevin Duh, and Paul McNamee. Identifying fluently inadequate output in neural and statistical machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 233–243, Dublin, Ireland, August 2019.

[165] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, jul 2004.

[166] Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. GLEU: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 344–351, Prague, Czech Republic, jun 2007.

[167] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[168] Mateusz Barański, Jan Jasiński, Julitta Bartolewska, Stanisław Kacprzak, Marcin Witkowski, and Konrad Kowalczyk. Investigation of whisper asr hallucinations induced by non-speech audio. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025.

[169] Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

[170] Frank Wessel, Ralf Schluter, Klaus Macherey, and Hermann Ney. Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on speech and audio processing*, 9(3):288–298, 2002.

[171] S. Cox and R. Rose. Confidence measures for the switchboard database. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 511–514 vol. 1, 1996.

[172] Thomas Kemp and Thomas Schaaf. Estimating confidence using word lattices. In *5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, pages 827–830, 1997.

[173] Iain McCowan, Darren Moore, John Dines, Daniel Gatica-Perez, Mike Flynn, Pierre Wellner, and Hervé Bourlard. On the use of information retrieval measures for speech recognition evaluation. *Idiap-RR Idiap-RR-73-2004, IDIAP, Martigny, Switzerland, 0*, 2004.

[174] Andrew Cameron Morris, Viktoria Maier, and Phil Green. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Interspeech 2004*, pages 2765–2768, 2004.

[175] Hanin Atwany, Abdul Waheed, Rita Singh, Monojit Choudhury, and Bhiksha Raj. Lost in transcription, found in distribution shift: Demystifying hallucination in speech foundation models. *arXiv preprint arXiv:2502.12414*, 2025.

[176] Arie Cattan, Paul Roit, Shiyue Zhang, David Wan, Roee Aharoni, Idan Szpektor, Mohit Bansal, and Ido Dagan. Localizing factual inconsistencies in attributable text generation. *arXiv preprint arXiv:2410.07473*, 2024.

[177] Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[178] Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. Fine-grained hallucination detection and editing for language models. In *First Conference on Language Modeling*, 2024.

[179] Prashant Serai, Vishal Sunder, and Eric Fosler-Lussier. Hallucination of speech recognition errors with sequence to sequence learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:890–900, 2022.

[180] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *CVPR*, 2020.

[181] Jamie Hayes, Ilia Shumailov, Eleni Triantafillou, Amr Khalifa, and Nicolas Papernot. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy. *arXiv preprint arXiv:2403.01218*, 2024.

[182] Matthew Jagielski, Om Thakkar, Florian Tramer, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, et al. Measuring forgetting of memorized training examples. *arXiv preprint arXiv:2207.00099*, 2022.

[183] Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. Muse: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460*, 2024.

[184] Dasol Choi and Dongbin Na. Towards machine unlearning benchmarks: Forgetting the personal identities in facial recognition systems. *arXiv preprint arXiv:2311.02240*, 2023.

[185] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*, 2024.

[186] Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in llms. *arXiv preprint arXiv:2310.02238*, 2023.

[187] Orchid Chetia Phukan, Girish, Mohd Mujtaba Akhtar, Shubham Singh, Swarup Ranjan Behera, Vandana Rajan, Muskaan Singh, Arun Balaji Buduru, and Rajesh Sharma. Towards Machine Unlearning for Paralinguistic Speech Processing. In *Interspeech 2025*, pages 4473–4477, 2025.

[188] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pages 141–159. IEEE, 2021.

[189] Woosung Choi, Junghyun Koo, Kin Wai Cheuk, Joan Serrà, Marco A Martínez-Ramírez, Yukara Ikemiya, Naoki Murata, Yuhta Takida, Wei-Hsiang Liao, and Yuki Mitsufuji. Large-scale training data attribution for music generative models via unlearning. *arXiv preprint arXiv:2506.18312*, 2025.

[190] Jiali Cheng and Hadi Amiri. Speech Unlearning. In *Interspeech 2025*, pages 3209–3213, 2025.

[191] Rodrigo Castellon, Chris Donahue, and Percy Liang. Codified audio language modeling learns useful representations for music information retrieval. In *ISMIR*, 2021.

[192] Tung-Yu Wu, Tsu-Yuan Hsu, Chen-An Li, Tzu-Han Lin, and Hung-yi Lee. The efficacy of self-supervised speech models for audio representations. In *PMLR*, 2022.

[193] Santiago Pascual, Mirco Ravanelli, Joan Serrà, Antonio Bonafonte, and Yoshua Bengio. Learning Problem-Agnostic Speech Representations from Multiple Self-Supervised Tasks. In *Proc. Interspeech*, 2019.

[194] Salah Zaiem, Titouan Parcollet, Slim Essid, and Abdelwahab Heba. Pretext tasks selection for multitask self-supervised audio representation learning. *IEEE J. Sel. Top. Signal Process.*, 2022.

[195] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. BYOL for Audio: Exploring pre-trained general-purpose audio representations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[196] Solène Evain and et al. Lebenchmark: A reproducible framework for assessing self-supervised representation learning from speech. In *Proc. Interspeech*, 2021.

[197] Luyu Wang and et a. Towards learning universal audio representations. In *ICASSP*, 2022.

[198] Sreyan Ghosh, Ashish Seth, and S Umesh. Decorrelating feature spaces for learning general-purpose audio representations. *IEEE J. Sel. Top. Signal Process.*, 2022.

[199] Joseph Turian and et al. Hear: Holistic evaluation of audio representations. In *NeurIPS Competitions and Demonstrations Track*. PMLR, 2022.

[200] Lawrence Philips. The double metaphone search algorithm. *Dr. Dobb's Journal*, June 2000. Available at https://drdobbs.com/the-double-metaphone-search-algorithm/184401251?pgno=2.

[201] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[202] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

[203] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 7871. Association for Computational Linguistics, 2020.

[204] François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20*, pages 198–208. Springer, 2018.

[205] Guoguo Chen et al. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *Interspeech 2021*, pages 3670–3674, 2021.

[206] Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, David Snyder, Aswin Shanmugam Subramanian, Jan Trmal, Bar Ben Yair, Christoph Boeddeker, Zhaoheng Ni, Yusuke Fujita, Shota Horiguchi, Naoyuki Kanda, Takuya Yoshioka, and Neville Ryant. Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. In *6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, pages 1–7, 2020.

[207] Tyler Kendall and Charlie Farrington. The corpus of regional african american language, 2023. Accessed via The Online Resources for African American Language Project.

[208] Wenbin Wang, Yang Song, and Sanjay Jha. Globe: A high-quality english corpus with global accents for zero-shot speaker adaptive text-to-speech, 2024.

[209] Junbo Zhang et al. speechocean762: An open-source non-native english speech corpus for pronunciation assessment. In *Proc. Interspeech 2021*, 2021.

[210] Sameer Pradhan, Ronald A. Cole, and Wayne H. Ward. My science tutor (MyST)–a large corpus of children‘s conversational speech. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12040–12045, Torino, Italia, May 2024. ELRA and ICCL.

[211] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024.

[212] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[213] Krishna C Puvvada, Piotr Żelasko, He Huang, Oleksii Hrinchuk, Nithin Rao Koluguri, Kunal Dhawan, Somshubra Majumdar, Elena Rastorgueva, Zhehuai Chen, Vitaly Lavrukhin, et al. Less is more: Accurate speech recognition & translation without web-scale data. *arXiv preprint arXiv:2406.19674*, 2024.

[214] IBM Granite Team. Granite 3.0 language models, 2024.

[215] Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

[216] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[217] Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure FP Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, et al. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *arXiv preprint arXiv:2310.00274*, 2023.

[218] Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. *NeurIPS*, 36, 2024.

[219] Shashwat Goel, Ameya Prabhu, Amartya Sanyal, Ser-Nam Lim, Philip Torr, and Ponnurangam Kumaraguru. Towards adversarial evaluations for inexact machine unlearning. *arXiv preprint arXiv:2201.06640*, 2022.

[220] Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7210–7217, 2023.

[221] Beomseok Lee, Ioan Calapodescu, Marco Gaido, Matteo Negri, and Laurent Besacier. Speech-massive: A multilingual speech dataset for slu and beyond. In *Proc. INTERSPEECH 2024*, pages 817–821, 2024.

[222] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *Interspeech*, 2021.

[223] Keltin Grimes, Collin Abidi, Cole Frank, and Shannon Gallagher. Gone but not forgotten: Improved benchmarks for machine unlearning. *arXiv preprint arXiv:2405.19211*, 2024.

[224] Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. Rwku: Benchmarking real-world knowledge unlearning for large language models. *arXiv preprint arXiv:2406.10890*, 2024.

[225] Karol J. Piczak. Esc: Dataset for environmental sound classification. In *ACM Multimedia*, MM '15, New York, NY, USA, 2015. ACM.

[226] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *ACM Multimedia*, MM '14, New York, NY, USA, 2014. ACM.

[227] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. Fsd50k: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.

[228] Natalie Holz, Pauline Larrouy-Maestri, and David Poeppel. The variably intense vocalizations of affect and emotion (vivae) corpus prompts new perspective on nonspeech perception. *Emotion*, 2022.

[229] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis. In *ISMIR*, 2017.

[230] Edith Law, Kris West, Michael I Mandel, Mert Bay, and J Stephen Downie. Evaluation of algorithms using games: The case of music tagging. In *ISMIR*, 2009.

[231] Juan J Bosch, Jordi Janer, Ferdinand Fuhrmann, and Perfecto Herrera. A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals. In *ISMIR*, 2012.

[232] Vincent Lostanlen and Carmine-Emanuele Cella. Deep convolutional networks on the pitch spiral for musical instrument recognition. In *ISMIR*, 2016.

[233] Sören Becker, Johanna Vielhaben, Marcel Ackermann, Klaus-Robert Müller, Sebastian Lapuschkin, and Wojciech Samek. AudioMNIST: Exploring explainable artificial intelligence for audio analysis on a simple benchmark. *Journal of the Franklin Institute*, 2024.

[234] Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco. EMOVO corpus: an Italian emotional speech database. In *LREC*. European Language Resources Association (ELRA), 2014.

[235] Alexei Baevski and et al. data2vec: A general framework for self-supervised learning in speech, vision and language. In *ICML*. PMLR, 2022.

[236] Jacob Kahn and et al. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP*. IEEE, 2020.

[237] Jort F. Gemmeke and et al. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017.

[238] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. In *Proc. Interspeech 2022*, pages 2753–2757, 2022.

[239] Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. Ssast: Self-supervised audio spectrogram transformer. In *AAAI*, 2022.

[240] Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D Manning. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, 2017.

[241] Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*, 2018.

[242] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[243] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. In Greg Kondrak and Taro Watanabe, editors, *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[244] Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3002–3017, Online, June 2021. Association for Computational Linguistics.

[245] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

[246] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. In *The Twelfth International Conference on Learning Representations*, 2024.

[247] Matthew Henderson, Blaise Thomson, and Jason D. Williams. The second dialog state tracking challenge. In Kallirroi Georgila, Matthew Stone, Helen Hastie, and Ani Nenkova, editors, *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272, Philadelphia, PA, U.S.A., June 2014. Association for Computational Linguistics.

[248] Koichiro Yoshino, Yun-Nung Chen, Paul Crook, Satwik Kottur, Jinchao Li, Behnam Hedayatnia, Seungwhan Moon, Zhengcong Fei, Zekang Li, Jinchao Zhang, et al. Overview of the tenth dialog system technology challenge: Dstc10. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:765–778, 2023.

[249] Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. *Advances in Neural Information Processing Systems*, 36:39088–39118, 2023.

[250] Keon Lee, Kyumin Park, and Daeyoung Kim. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[251] Se Jin Park, Chae Won Kim, Hyeongseop Rha, Minsu Kim, Joanna Hong, Jeong Hun Yeo, and Yong Man Ro. Let's go real talk: Spoken dialogue model for face-to-face conversation. In *The 62nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2024.

[252] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019.

[253] Tu Anh Nguyen, Wei-Ning Hsu, Antony d'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, et al. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. In *INTERSPEECH 2023-24th Annual Conference of the International Speech Communication Association*, pages 4823–4827. ISCA, 2023.

[254] Guan-Ting Lin, Cheng-Han Chiang, and Hung-Yi Lee. Advancing large language models to capture varied speaking styles and respond properly in spoken conversations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6626–6642, 2024.

[255] Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[256] Hongfei Xue, Yuhao Liang, Bingshen Mu, Shiliang Zhang, Mengzhe Chen, Qian Chen, and Lei Xie. E-chat: Emotion-sensitive spoken dialogue system with large language models. In *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 586–590. IEEE, 2024.

[257] Panagiotis Tzirakis, Jiehao Zhang, and Bjorn W Schuller. End-to-end speech emotion recognition using deep neural networks. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5089–5093. IEEE, 2018.

[258] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462, New York, NY, USA, 2010. Association for Computing Machinery.

[259] Alan Cowen, Disa Sauter, Jessica L Tracy, and Dacher Keltner. Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression. *Psychological Science in the Public Interest*, 20(1):69–90, 2019.

[260] Lorenzo Vaiani, Alkis Koudounas, Moreno La Quatra, Luca Cagliero, Paolo Garza, and Elena Baralis. Transformer-based non-verbal emotion recognition: Exploring model portability across speakers' genders. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, MuSe' 22, page 89–94, New York, NY, USA, 2022. Association for Computing Machinery.

[261] Reshmashree B Kantharaju, Fabien Ringeval, and Laurent Besacier. Automatic recognition of affective laughter in spontaneous dyadic interactions from audiovisual signals. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 220–228, 2018.

[262] Scott Condron, Georgia Clarke, Anita Klementiev, Daniela Morse-Kopp, Jack Parry, and Dimitri Palaz. Non-verbal vocalisation and laughter detection using sequence-to-sequence models and multi-label training. In *Interspeech*, pages 2506–2510, 2021.

[263] Panagiotis Tzirakis, Alice Baird, Jeffrey Brooks, Christopher Gagne, Lauren Kim, Michael Opara, Christopher Gregory, Jacob Metrick, Garrett Boseck, Vineet Tiruvadi, et al. Large-scale nonverbal vocalization detection using transformers. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[264] Detai Xin, Shinnosuke Takamichi, and Hiroshi Saruwatari. Exploring the effectiveness of self-supervised learning and classifier chains in emotion recognition of nonverbal vocalizations. *arXiv preprint arXiv:2206.10695*, 2022.

[265] Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages, 2022.

[266] Valentina Bellomaria, Giuseppe Castellucci, Andrea Favalli, and Raniero Romagnoli. Almawave-slu: A new dataset for SLU in italian. *CoRR*, abs/1907.07526, 2019.

[267] Moreno La Quatra and Luca Cagliero. Bart-it: An efficient sequence-to-sequence model for italian text summarization. *Future Internet*, 15(1), 2023.

[268] Gabriele Sarti and Malvina Nissim. It5: Large-scale text-to-text pretraining for italian language understanding and generation. *arXiv preprint arXiv:2203.03759*, 2022.

[269] Pierpaolo Basile, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. Llamantino: Llama 2 models for effective text generation in italian language. *arXiv preprint arXiv:2312.09993*, 2023.

[270] Alan Ramponi and Camilla Casula. DiatopIt: A corpus of social media posts for the study of diatopic language variation in Italy. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 187–199, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[271] Alan Ramponi and Camilla Casula. Geolingit at evalita 2023: Overview of the geolocation of linguistic variation in italy task. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR. org, Parma, Italy*, 2023.

[272] Marco Marini, Mauro Viganò, Massimo Corbo, Marina Zettin, Gloria Simoncini, Bruno Fattori, Clelia D'Anna, Massimiliano Donati, and Luca Fanucci. Idea: An italian dysarthric speech database. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 1086–1093, 2021.

[273] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805, 2022.

[274] Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Interpreting and explaining deep neural networks for classification of audio signals. *arXiv preprint arXiv:1807.03418*, 2018.

[275] Annika Frommholz, Fabian Seipel, Sebastian Lapuschkin, Wojciech Samek, and Johanna Vielhaben. Xai-based comparison of input representations for audio event classification. *arXiv preprint arXiv:2304.14019*, 2023.

[276] Xiaoliang Wu, Peter Bell, and Ajitha Rajan. Explanations for automatic speech recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[277] Saumitra Mishra, Bob L Sturm, and Simon Dixon. Local interpretable model-agnostic explanations for music content analysis. In *ISMIR*, volume 53, pages 537–543, 2017.

[278] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[279] Xiaoliang Wu, Peter Bell, and Ajitha Rajan. Can we trust explainable ai methods on asr? an evaluation on phoneme recognition. *arXiv preprint arXiv:2305.18011*, 2023.

[280] Soroosh Mariooryad and Carlos Busso. Generating human-like behaviors using joint, speech-driven models for conversational agents. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(8):2329–2340, 2012.

[281] Tiantian Feng and Shrikanth Narayanan. Foundation model assisted automatic speech emotion recognition: Transcribing, annotating, and augmenting. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12116–12120, 2024.

[282] Eduardo Fonseca, Jordi Pons Puig, Xavier Favory, Frederic Font Corbera, Dmitry Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. Freesound datasets: a platform for the creation of open audio datasets. In *Hu X, Cunningham SJ, Turnbull D, Duan Z, editors. Proceedings of the 18th ISMIR Conference; 2017 oct 23-27; Suzhou, China.[Canada]: International Society for Music Information Retrieval; 2017. p. 486-93.* International Society for Music Information Retrieval (ISMIR), 2017.

[283] Huang-Cheng Chou, Wei-Cheng Lin, Lien-Chiang Chang, Chyi-Chang Li, Hsi-Pin Ma, and Chi-Chun Lee. Nnime: The nthu-ntua chinese interactive multimodal emotion corpus. In *2017 Seventh international conference on affective computing and intelligent interaction (ACII)*, pages 292–298. IEEE, 2017.

[284] Muhammad Mamunur Rashid, Guiqing Li, and Chengrui Du. Nonspeech7k dataset: Classification and analysis of human non-speech sound. *IET Signal Processing*, 17(6):e12233, 2023.

[285] Kristina T Johnson, Jaya Narain, Thomas Quatieri, Pattie Maes, and Rosalind W Picard. Recanvo: A database of real-world communicative and affective nonverbal vocalizations. *Scientific Data*, 10(1):523, 2023.

[286] Dawn AA Black, Ma Li, and Mi Tian. Automatic identification of emotional cues in chinese opera singing. *ICMPC, Seoul, South Korea*, 2014.

[287] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Tut database for acoustic scene classification and sound event detection. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 1128–1132. IEEE, 2016.

[288] Mark Cartwright and Bryan Pardo. Vocalsketch: Vocally imitating audio concepts. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 43–46, 2015.

[289] Yuan Gong, Jin Yu, and James Glass. Vocalsound: A dataset for improving human vocal sounds recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 151–155. IEEE, 2022.

[290] Dejoli Landry, Qianhua He, Haikang Yan, and Yanxiong Li. Asvp-esd: A dataset and its benchmark for emotion recognition using both speech and non-speech utterances. *Global Scientific Journals*, 8:1793–1798, 2020.

[291] Scott Condron, Georgia Clarke, Anita Klementiev, Daniela Morse-Kopp, Jack Parry, and Dimitri Palaz. Non-verbal vocalisation and laughter detection using sequence-to-sequence models and multi-label training. In *Interspeech*, 2023.

[292] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*, 2023.

[293] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[294] Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350, 2001.

[295] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[296] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[297] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[298] Team Cohere, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, et al. Command a: An enterprise-ready large language model. *arXiv preprint arXiv:2504.00698*, 2025.

[299] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[300] Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87, 2023.

[301] Joseph L Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619, 1973.

[302] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[303] OpenAI. Gpt-4o-mini: Advancing cost-efficient intelligence, 2024.

[304] Google. Gemini 2.0 flash, 2025.

[305] Google. Gemini 2.5 flash preview, 2025.

[306] Google. Gemini 2.5 pro preview, 2025.

[307] Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. Xtts: a massively multilingual zero-shot text-to-speech model. In *Interspeech 2024*, pages 4978–4982, 2024.

[308] Canopy Labs. Orpheus-3b-0.1-ft: A multilingual text-to-speech model. https://huggingface.co/canopylabs/orpheus-3b-0.1-ft, 2025. Fine-tuned version of Orpheus for expressive TTS.

[309] Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation, 2025.

[310] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911, 2014.

[311] Martin Maiden and Mair Parry. *The dialects of Italy*. Routledge, 2006.

[312] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

[313] Stefan Schweter. Italian bert and electra models, November 2020.

[314] Richard P. Lippmann. Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15, 1997.

[315] Francesco Avolio. *Lingue e dialetti d'Italia*, volume 380. Carocci Roma, Italy, 2009.

[316] Alan Ramponi. Nlp for language varieties of italy: Challenges and the path forward. *arXiv preprint arXiv:2209.09757*, 2022.

[317] Gaetano Berruto et al. Dialect/standard convergence, mixing, and models of language contact: the case of italy. *Dialect change. Convergence and divergence in European languages*, pages 81–97, 2005.

[318] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[319] Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Bmvc*, volume 1, page 3, 2016.

[320] Michele Castellarin and Fabio Tosques. Das vivaio acustico delle lingue e dei dialetti d'italia (vivaldi): Ein sprachatlas als nützliches tool für die untersuchung italienischer dialekte und minderheitensprachen. *kunsttexte. de-Journal für Kunst-und Bildgeschichte*, 2:1–14, 2013.

[321] Roland Bauer. Vivaldi-sicilia. documentazione sonora dei dialetti siciliani. In *Percorsi di Geografia linguistica. Idee per un atlante siciliano della cultura dialettale e dell'italiano regionale*, pages 543–550. Centro di Studi Filologici e Linguistici Siciliani/Istituto di Filologia e ..., 1995.

[322] Dieter Kattenbusch and Carola Köhler. La sardegna nel progetto vivaldi. *GRIMALDI, LUCIA & MENSCHING, GUIDO (a cura di), Su sardu. Limba de Sardigna e limba de Europa. Cagliari: Cuec*, pages 193–203, 2004.

[323] Dieter Kattenbusch, Fabio Tosques, and Andreas Rauher. „umbria dialettale ". *Claudia Schlaak, Lena Busse (Hg.), Sprachkontakte, Sprachvariation und Sprachwandel, Tübingen*, pages 443–460, 2011.

[324] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Proc. Interspeech 2020*, pages 3830–3834, 2020.

[325] Jörgen Valk and Tanel Alumäe. Voxlingua107: a dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658. IEEE, 2021.

[326] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[327] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[328] RR Selvaraju, M Cogswell, A Das, R Vedantam, D Parikh, and D Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *arXiv preprint arXiv:1610.02391*, 2022.

[329] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*, 2023.

[330] Ian C Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *The Journal of Machine Learning Research*, 22(1):9477–9566, 2021.

[331] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[332] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*, 2020.

[333] Juan Rafael Orozco-Arroyave, Julián David Arias-Londoño, Jesús Francisco Vargas-Bonilla, María Claudia Gonzalez-Rátiva, and Elmar Nöth. New spanish speech corpus database for the analysis of people suffering from parkinson's disease. In *Lrec*, pages 342–347, 2014.

[334] Moreno La Quatra, Maria Francesca Turco, Torbjørn Svendsen, Giampiero Salvi, Juan Rafael Orozco-Arroyave, and Sabato Marco Siniscalchi. Exploiting foundation models and speech enhancement for parkinson's disease detection from speech in real-world operative conditions. *Interspeech 2024*, 2024.

[335] Davide Ghia, Gabriele Ciravegna, Alkis Koudounas, Marco Fantini, Erika Crosetti, Giovanni Succo, and Tania Cerquitelli. A concept-based approach to voice disorder detection. *arXiv preprint arXiv:2507.17799*, 2025.

[336] Nelson Roy, Ray M Merrill, Steven D Gray, and Elaine M Smith. Voice disorders in the general population: prevalence, risk factors, and occupational impact. *The Laryngoscope*, 2005.

[337] Seth M Cohen. Self-reported impact of dysphonia in a primary care population: An epidemiological study. *The Laryngoscope*, 2010.

[338] Neil Bhattacharyya. The prevalence of voice problems among adults in the united states. *The Laryngoscope*, 2014.

[339] Nikolaos Spantideas, Eirini Drosou, Anna Karatsis, and Dimitrios Assimakopoulos. Voice disorders in the general greek population and in patients with laryngopharyngeal reflux. prevalence and risk factors. *Journal of Voice*, 2015.

[340] Marco Fantini, Gabriele Ciravegna, Alkis Koudounas, Tania Cerquitelli, Elena Baralis, Giovanni Succo, and Erika Crosetti. The rapidly evolving scenario of acoustic voice analysis in otolaryngology. *Cureus*, 16(11), 2024.

[341] Gabriele Ciravegna, Alkis Koudounas, Marco Fantini, Tania Cerquitelli, Elena Baralis, Erika Crosetti, Giovanni Succo, et al. Non-invasive ai-powered diagnostics: The case of voice-disorder detection-vision paper. In *Proceedings of the Workshops of the EDBT/ICDT 2024 Joint Conference*, volume 3651. CEUR, 2024.

[342] Qingqing Liu, Gabriele Ciravegna, Alkis Koudounas, Tania Cerquitelli, Elena Baralis, et al. Multimodal fusion techniques to enhance voice disorder diagnoses. In *CEUR WORKSHOP PROCEEDINGS*, volume 3946. CEUR, 2025.

[343] Elke Brunner, Katharina Eberhard, and Markus Gugatschka. Prevalence of benign vocal fold lesions: Long-term results from a single european institution. *Journal of Voice*, 2023.

[344] Ibrahim Karabayir, Samuel M Goldman, Suguna Pappu, and Oguz Akbilgic. Gradient boosting for parkinson's disease diagnosis from voice recordings. *BMC Medical Informatics and Decision Making*, 2020.

[345] Helder Vieira, Nelson Costa, Tomás Sousa, Sara Reis, and Luis Coelho. Voice-based classification of amyotrophic lateral sclerosis: where are we and where are we going? a systematic review. *Neurodegenerative Diseases*, 2020.

[346] Wing-Zin Leung, Mattias Cross, Anton Ragni, and Stefan Goetze. Training data augmentation for dysarthric automatic speech recognition by text-to-dysarthric-speech synthesis. In *Interspeech*, 2024.

[347] Dhvani Shah, Vanshika Lal, Zihan Zhong, Qianli Wang, and Seyed Reza Shahamiri. Dysarthric speech recognition: A comparative study. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 89–94, 2023.

[348] Mark Hasegawa-Johnson et al. Community-supported shared infrastructure in support of speech accessibility. *Journal of Speech, Language, and Hearing Research*, 2024.

[349] Lotfi Salhi, Mourad Talbi, and Adnane Cherif. Voice disorders identification using hybrid approach: Wavelet analysis and multilayer neural networks. *International Journal of Electrical and Computer Engineering*, 2(9):3003–3012, 2008.

[350] Julián David Arias-Londoño, Jorge Andrés Gómez-García, Laureano Moro-Velázquez, and Juan Ignacio Godino-Llorente. Byovoz automatic voice condition analysis system for the 2018 femh challenge. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018.

[351] Xiangyu Peng, Huoyao Xu, Jie Liu, Junlang Wang, and Chaoming He. Voice disorder classification using convolutional neural network based on deep transfer learning. *Scientific Reports*, 13(1):7264, 2023.

[352] Xiaoping Xie, Hao Cai, Can Li, Yu Wu, and Fei Ding. A voice disease detection method based on mfccs and shallow cnn. *Journal of Voice*, 2023.

[353] Umesh Kumar Lilhore, Surjeet Dalal, Neetu Faujdar, Martin Margala, Prasun Chakrabarti, Tulika Chakrabarti, Sarita Simaiya, Pawan Kumar, Pugazhenthan Thangaraju, and Hemasri Velmurugan. Hybrid cnn-lstm model with efficient hyperparameter tuning for prediction of parkinson's disease. *Scientific Reports*, 13(1):14605, 2023.

[354] Rumana Islam, Esam Abdel-Raheem, and Mohammed Tarique. Voice pathology detection using convolutional neural networks with electroglottographic (egg) and speech signals. *Computer Methods and Programs in Biomedicine Update*, 2:100074, 2022.

[355] Seyed Reza Shahamiri, Vanshika Lal, and Dhvani Shah. Dysarthric speech transformer: A sequence-to-sequence dysarthric speech recognition system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

[356] Ahmad Almadhor, Rizwana Irfan, Jiechao Gao, Nasir Saleem, Hafiz Tayyab Rauf, and Seifedine Kadry. E2e-dasr: End-to-end deep learning-based dysarthric automatic speech recognition. *Expert Systems with Applications*, 222:119797, 2023.

[357] Loukas Ilias, Dimitris Askounis, and John Psarras. Detecting dementia from speech and transcripts using transformers. *Computer Speech & Language*, 79:101485, 2023.

[358] Shao-Hsuan Lee, Jen-Fang Yu, Tuan-Jen Fang, and Guo-She Lee. Vocal fold nodules: A disorder of phonation organs or auditory feedback? *Clinical Otolaryngology*, 2019.

[359] Gaston Schlotthauer, María Eugenia Torres, and María Cristina Jackson-Menaldi. A pattern recognition approach to spasmodic dysphonia and muscle tension dysphonia automatic classification. *Journal of voice*, 2010.

[360] Huimeng Wang, Zengrui Jin, Mengzhe Geng, Shujie Hu, Guinan Li, Tianzi Wang, Haoning Xu, and Xunying Liu. Enhancing pre-trained asr system fine-tuning for dysarthric speech recognition using adversarial data augmentation. In *ICASSP*, 2024.

[361] I-Ting Hsieh and Chung-Hsien Wu. Dysarthric speech recognition using curriculum learning and articulatory feature embedding. In *Interspeech 2024*, pages 1300–1304, 2024.

[362] Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Eng-Siong Chng. Hyporadise: An open baseline for generative speech recognition with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[363] Moreno La Quatra, Valerio Mario Salerno, Yu Tsao, and Sabato Marco Siniscalchi. Flanec: Exploring flan-t5 for post-asr error correction. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 608–615. IEEE, 2024.

[364] Bogdan Woldert-Jokisz. Saarbruecken voice database. *Essen University Hospital*, 2007.

[365] Luis MT Jesus, Inês Belo, Jessica Machado, and Andreia Hall. The advanced voice function assessment databases (avfad): Tools for voice clinicians and speech research. In *Advances in Speech-language Pathology*. IntechOpen, 2017.

[366] Gail B Kempster, Bruce R Gerratt, Katherine Verdolini Abbott, Julie Barkmeier-Kraemer, and Robert E Hillman. Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol. *American Journal of Speech-Language Pathology*, 18:124–132, 2009.

[367] Moreno La Quatra, Nicole Dalia Cilia, Vincenzo Conti, Salvatore Sorce, Giovanni Garraffa, and Valerio Mario Salerno. Vision-language multimodal fusion in dermatological disease classification. In *Pattern Recognition. ICPR 2024 International Workshops and Challenges*, pages 211–225, Cham, 2025. Springer Nature Switzerland.

[368] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

[369] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

[370] Hyung Won Chung et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[371] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language resources and evaluation*, 46:523–541, 2012.

[372] Mario Zusag, Laurin Wagner, and Bernhad Thallinger. Crisperwhisper: Accurate timestamps on verbatim speech transcriptions. *Interspeech 2024*, 2024.

[373] David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015.

[374] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech*, 2019.

[375] Bornali Phukon, Xiuwen Zheng, and Mark Hasegawa-Johnson. Aligning ASR Evaluation with Human and LLM Judgments: Intelligibility Metrics Using Phonetic, Semantic, and NLI Approaches. In *Interspeech 2025*, pages 5708–5712, 2025.