Toward Robust, Responsible and Trustworthy Speech Foundation Models

Candidate: Alkis Koudounas
Supervisor: Prof. Elena Baralis
Co-supervisor: Prof. Eliana Pastor

Politecnico di Torino
Doctoral Program in Computer and Control Engineering (38th cycle)

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 hallucinations 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.