Transfer Learning of Large Speech Models for Italian Speech Emotion Recognition

Federico D'Asaro
Dip. di Automatica e Informatica
Politecnico di Torino
Torino (TO), Italy
federico.dasaro@polito.it

Juan José Márquez Villacís
AI, Data & Space (ADS)
LINKS Foundation
Torino (TO), Italy
juan.marquez@linksfoundation.com

Giuseppe Rizzo
AI, Data & Space (ADS)
LINKS Foundation
Torino (TO), Italy
giuseppe.rizzo@linksfoundation.com

Andrea Bottino
Dip. di Automatica e Informatica
Politecnico di Torino
Torino (TO), Italy
andrea.bottino@polito.it

***Abstract*—Recent research in Automated Speech Recognition has shifted towards using large pre-trained speech models trained on extensive corpora with a Self-Supervised Learning (SSL) approach. These models can transfer general-purpose knowledge to tasks like Speech Emotion Recognition (SER). Due to their highly parameterized architecture, fine-tuning all the weights is computationally inefficient. Consequently, new Parameter-Efficient Fine-Tuning (PEFT) strategies have been explored for the SER task in English. Given the lack of SSL speech models in Italian, current models are either English-only or multilingual, with little effort made to adapt them to SER in Italian. In this work, we investigate transfer learning performance on Italian SER using PEFT strategies, marking the first exploration in this direction. We apply PEFT techniques, such as Low-Rank Adaptation (LoRA) and Adapter, on Italian SER datasets *Emozionalmente*, *DEMoS*, and *EMOVO*. Results show LoRA is the most effective PEFT technique for Italian SER. Speech models pre-trained on large-scale English corpora perform comparably to, or better than, multilingual ones, even when specialized in Italian before the SER task, suggesting some shared paralinguistic features between the languages.**

***Index Terms*—Italian Speech Emotion Recognition, Parameter-efficient fine-tuning**

I. Introduction

Recognizing emotions from speech is crucial in human-computer interaction [1]. Speech Emotion Recognition (SER) classifies audio into discrete emotions [2] or within a continuous space [3]. Traditional SER methods extract hand-crafted features like prosodic [4], voice quality [5], and spectral features such as MFCC [6], which are then used by deep learning models for classification. Deep learning methods have evolved from Convolutional Neural Networks (CNNs) [7] and Recurrent Neural Networks (RNNs) [8] to Transformers [9], which are now used in large speech models pre-trained on large audio datasets via Self-Supervised Learning (SSL) techniques like masked speech modeling [10]. These models retain general-purpose knowledge and have shown remarkable results when fine-tuned for SER in English [11]–[14]. To improve transfer learning efficiency for SER in English, researchers [15], [16] have adopted Parameter Efficient Fine-Tuning (PEFT) [17] for models like Wav2Vec 2.0, WavLM [18], and Whisper [19]. PEFT adds task-specific components without altering pre-trained model parameters, avoiding the need to store multiple copies of large speech models for different tasks.

In this work, we further extend the investigation of transfer learning methodologies of large speech models for the task of cross-lingual speech emotion recognition in the Italian language. Most recent works still adopt traditional approaches that extract hand-crafted features and then use a CNN+MLP to obtain a suitable classifier [20]. A step towards using SSL has been made by [21], who fine-tuned the entire Wav2Vec 2.0 transformer for Italian SER. In contrast, we explore the use of PEFT strategies—Low-Rank Adaptation (LoRA) [22] and Adapter [17]—to adapt large speech models for Italian SER. Due to the scarcity of large speech models pre-trained on Italian, we utilize a variety of English-only and multilingual pre-trained speech models (Wav2Vec2.0, WavLM, Whisper, XLSR-53) and test them on Italian SER datasets *EMOVO*, *Emozionalmente*, and *DEMoS*. In addition to assessing the effectiveness of PEFT techniques, we examine the extent to which English-only pre-trained speech models can transfer knowledge to Italian SER without specializing in Italian, by comparing their performance with multilingual speech models. The main contributions of our work are:

- We investigate PEFT techniques in adapting large pre-trained speech models for cross-lingual Speech Emotion Recognition in the Italian language, showing that LoRA consistently delivers better performance across all pre-trained speech models.
- We compare the transfer learning performance of English-

only and multilingual speech models for Italian SER, showing that the former perform comparably or even better than the latter. This suggests that some paralinguistic features useful for Italian SER tasks are embedded in English-only speech models.

The following sections are organized as follows: Section II introduces the pre-trained speech models and the PEFT methods adopted in this article. Section III discusses the fine-tuning approaches and the structure of the downstream model classifier. Section IV details the experimental setup and results obtained. Finally, Section V presents conclusions and considerations for future investigations.

II. Related Work

A. Italian Speech Emotion Recognition

Early approaches to addressing the SER task in the Italian language are primarily based on traditional machine learning algorithms combined with fundamental hand-crafted features such as MFCC, PLPs (Perceptual Linear Predictive), and EMLBs (Mel Bank Spectrum) [23]. [24] explores the use of spectral features by applying support vector machines to MFMC (Mel Frequency Magnitude Coefficient) features, showcasing the robustness of discriminative models in distinguishing between different emotional states. [25] models the Italian SER task as a regression problem, employing a support vector regressor to predict valence and arousal on a continuous scale in a two-dimensional domain. With the advent of deep learning, researchers began utilizing CNNs [26] and RNNs to capture more complex patterns in speech data. [27] integrates CNNs for extracting features from spectrograms with RNNs for temporal sequence modeling. [28] introduce a cross-modal distillation approach to train smaller CNN-based speech models by utilizing a larger teacher model trained on facial expression datasets.

Advancements in Self-Supervised Learning have produced large pre-trained speech models that can be fine-tuned for Italian SER tasks. The limited Italian speech corpus necessitates using English-only or multilingual models. For instance, [21] shows the effectiveness of the multilingual XLSR-53 [29] for Italian SER. Our study evaluates more speech models and fine-tuning strategies to transfer knowledge from English-only and multilingual models to Italian SER.

B. Large Speech Models

Transformers [9] are foundational in creating pre-trained models in NLP [30] and Computer Vision [31]. Trained on large unlabeled corpora using Self-Supervised Learning (SSL), these models transfer general-purpose knowledge to various downstream tasks. Similarly, in Automated Speech Recognition (ASR), new architectures trained on extensive corpora like LibriSpeech [32], GigaSpeech [33], VoxPopuli [34], and Common Voice [35] leverage techniques like masked speech modeling [10]. This article explores their potential for transferring knowledge to Italian SER, focusing on the following widely-used models: Wav2Vec 2.0 [10], is the first successful SSL approach in ASR, using Transformers to learn discrete speech units via a quantization module.

Language	**Pre-trained Architecture**	**Params**	**Fine-tuning Scenario**
English	Wav2Vec 2.0 Base	95.04M	A
English	WavLM Base+	94.70M	A
English	Wav2Vec 2.0 Large	317M	A
			B
Multilingual	Whisper Small	88.15M	A
			B
Multilingual	XLSR-53	317M	A
			B

TABLE I
Summary of pre-trained speech backbones used in this work on the two scenarios (A) and (B).

XLSR-53 [29], extends Wav2Vec 2.0 by training on 53 languages, learning shared latent speech representations and performing well against language-specific models.

WavLM [18], building upon [36], introduces utterance mixing and gated relative position bias to better model spoken content while maintaining speaker identity, outperforming other large speech models.

Whisper [19], is an encoder-decoder architecture trained on a large multilingual corpus, excelling in recognizing accents and technical terminology. Starting from the log-Mel spectrogram, it predicts text captions and performs tasks like language identification, transcription, and translation.

Fine-tuning WavLM and Whisper for Italian SER remains underexplored. This study evaluates English-only and multilingual models under two scenarios: (A) direct fine-tuning on Italian SER, and (B) incorporating a self-supervised fine-tuning step on Italian corpora before the downstream task fine-tuning, as detailed in Section III-B. Table I summarizes the pre-trained models and fine-tuning scenarios.

C. PEFT Methods

Fine-tuning large speech models from scratch is computationally expensive. Parameter Efficient Fine-Tuning (PEFT) techniques [17] adapt these models for downstream tasks with minimal task-specific parameters. According to [37], PEFT techniques include *Additive Tuning*, which freezes the large model's parameters and adjusts only newly introduced parameters (e.g., Adapter [17], Parallel Adapter [38], and Prompt Tuning [39]), and *Reparameterization*, which utilizes low-rank transformations of the model's weight matrices, such as LoRA [22] and DoRA [40]. These techniques have proven effective in English SER [11], [13], particularly with strong performance from LoRA and Adapter. We apply these techniques to cross-lingual Italian SER.

III. Method

In this section, we detail our methodology to address the following: (i) Evaluating the effectiveness of PEFT techniques, specifically LoRA and Adapter, in transferring learning to Italian SER. (ii) Comparing the performance of English-only

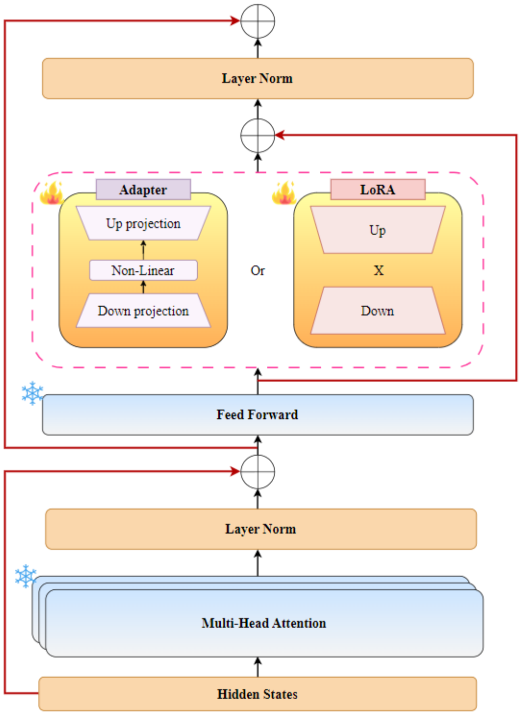

Fig. 1. System architecture incorporating Adapter and LoRA PEFT strategies within a Transformer block.

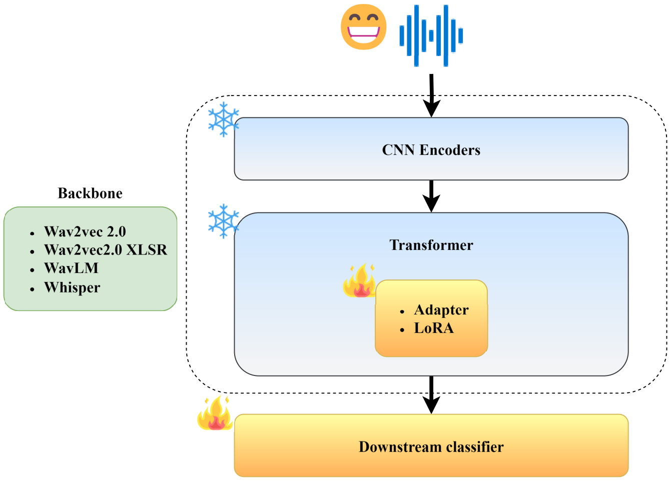

Fig. 2. Modeling framework used in this work. Alongside the downstream classifier which serves as a baseline, the PEFT methods are applied solely within the transformer blocks of the speech models.

and multilingual pre-trained speech models in transferring knowledge to Italian SER.

A. PEFT Integration

Since all evaluated speech models share a common structure of CNN encoders with Transformer blocks, our approach involves applying PEFT methods exclusively to the Transformer blocks while keeping the CNN parameters frozen, as illustrated in Figure 2. We describe the PEFT strategies used and the configuration of the downstream classifier integrated into the speech models.

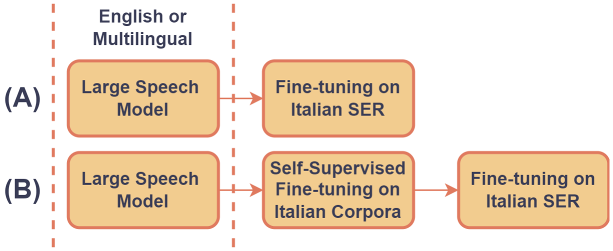
English or Multilingual
(A)
Large Speech Model
Fine-tuning on Italian SER
(B)
Large Speech Model
Self-Supervised Fine-tuning on Italian Corpora
Fine-tuning on Italian SER

Fig. 3. Fine-tuning scenarios evaluated in this work: (A) The speech model is directly fine-tuned on the Italian SER task. (B) The speech model is first fine-tuned using an SSL approach on a large Italian audio corpus before being fine-tuned on the Italian SER task.

We integrate Adapter and LoRA methods into each pre-trained speech model. For Adapter and LoRA, only their respective parameters are fine-tuned, while all other parameters of the speech model remain frozen. Additionally, following the method in [15], we implement a baseline fine-tuning approach where the entire speech model is frozen, and only a downstream classifier is fine-tuned. The same downstream classification architecture is used across all fine-tuning methods.

Our downstream model follows the architecture outlined in [15]. It takes the average pooling of all encoder hidden states as input. Each pooled hidden state passes through a feed-forward network with ReLU activations, followed by averaging over the temporal dimension before classification.

B. Specialization of Speech Models on Italian

In addition to PEFT techniques, we compare the performance of English-only and multilingual speech models. Specifically, in assessing their effectiveness, we consider two scenarios: (A) direct fine-tuning of the large speech model on Italian SER following the PEFT integration detailed in Sec. III-A, and (B) initial fine-tuning of the large speech model using a self-supervised learning approach with masked speech modeling for ASR tasks, incorporating at least 200 hours of Italian speech data, followed by fine-tuning on Italian SER (see Figure 3). This latter SSL approach has previously been implemented using the Common Voice dataset [21] with XLSR-53 for Italian SER. In this study, we expand this evaluation to include multilingual Whisper Small and English-only Wav2Vec 2.0. In this manner, we enrich the comparison between English-only and multilingual models by investigating whether additional specialization in the Italian language improves their ability to handle language nuances effectively in Italian during SER tasks.

IV. Experiments

A. Datasets and Metrics

To test our approaches and validate our observations, we conduct extensive experiments on three available datasets for Italian SER: EMOVO [41], Emozionalmente [21], and DEMoS

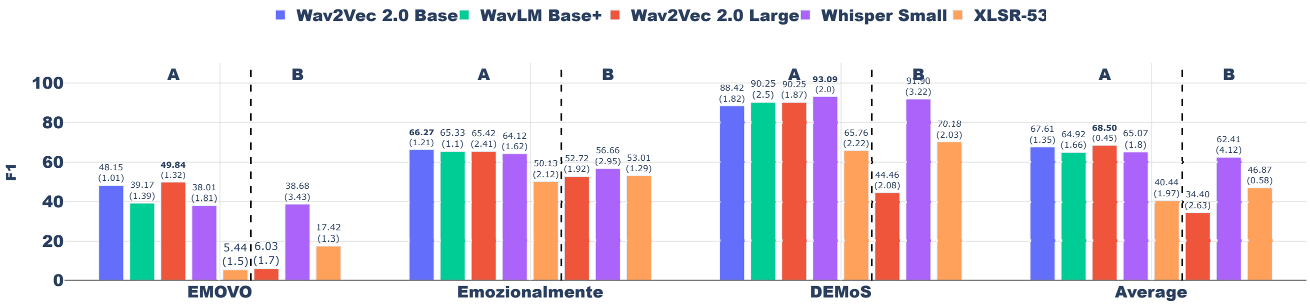

Fig. 4. Performance with fine-tuning downstream classification model (with the pre-trained backbone frozen) for SER in the two scenarios (A) and (B) as defined in Figure 3. Values are presented as the mean (standard deviation) across three runs.

Datasets	Anger	Disgust	Fear	Joy	Neutrality	Sadness	Surprise	Total
Emovo	84	84	84	84	84	84	84	588
Emozionalmente	986	986	986	986	986	986	986	6902
Demos	1477	1678	1156	1395	332	1530	1000	8568

TABLE II
SUMMARY OF DATASET STATISTICS USED IN THIS WORK.

[42]. Each dataset consists of audio recordings labeled with one of the six basic emotions commonly referred to as The Big Six [2]: anger, disgust, fear, joy, sadness, and surprise, plus a neutral state. Summary statistics of the datasets are presented in Table II.

EMOVO comprises 588 audio samples from 6 actors (3 males and 3 females). Each actor recorded 14 sentences, each simulating one of the Big Six emotions plus a neutral state. The recordings were recorded using professional equipment at a sample rate of 48 kHz, 16-bit stereo, in wav format.

Emozionalmente includes 6902 audio samples from 431 amateur actors (131 males, 299 females, and 1 identified as ”other”). Each actor verbalized 18 different sentences expressing the Big Six emotions plus neutrality. The recordings, with an average duration of 3.81 seconds, were made using non-professional equipment at a sample rate of 48 kHz, 16-bit mono, and stored in wav format.

DEMoS is the largest Italian SER dataset to date, consisting of 9697 samples collected from 68 voluntary students: 23 females, representing almost 35% of the samples, and 45 males. DEMoS includes also a secondary emotion, guilt, which we exclude to maintain consistent labeling across the three datasets used. The audio recordings have an average duration of 2.9 seconds and were recorded at 48 kHz, 16-bit mono, in wav format.

For quantitative evaluation, we use the F1 score, defined as:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

where Precision $= \frac{TP}{TP+FP}$ and Recall $= \frac{TP}{TP+FN}$, with TP representing true positives, FP representing false positives, and FN representing false negatives.

B. Experimental Details

The evaluation is conducted in a speaker-independent setting, with the data split into stratified train/validation/test sets in percentages of 80/10/10. To ensure that an actor is not present in multiple splits, we first group audios by actor and then apply the stratified split. We resample audios to 16kHz and truncate them to a maximum duration of 4 seconds [21]. We repeat the training three times with different seeds and average the results.

We adopt pre-trained Wav2Vec 2.0 Base (A)[1], WavLM Base+ (A)[2], Whisper Small (A)[3], Wav2Vec 2.0 Large (A)[4] and XLSR-53 (A)[5] checkpoints from Hugging Face to run experiments on scenario (A). For scenario (B), we adopt Whisper Small (B)[6], Wav2Vec 2.0 Large (B)[7] and XLSR-53 (B)[8], which have been fine-tuned on Italian using the Common Voice dataset[9].

We fine-tune the speech models for 30 epochs using the Adam optimizer with the following parameters: betas set to (0.9, 0.98) and epsilon to 1.0e-08. The learning rate is set to 5.0e-04 with a weight decay of 1.0e-04, following an exponential schedule. Following [15], we set the Adapter's hidden state dimension to 128, the LoRA rank to 8, and the classifier projection to 256.

Overall, we fine-tune every model on each dataset using the three PEFT strategies, resulting in a total of 72 training sessions.

C. Results

1) Downstream Classifier Performance: We begin by comparing the performance of the downstream models while freezing the pre-trained backbones, which serve as a baseline for the subsequent PEFT strategies (Figure 4). Starting with models in scenario (A), we observe that, on average,

[1] facebook/wav2vec2-base
[2] microsoft/wavlm-base-plus
[3] openai/whisper-small
[4] facebook/wav2vec2-large
[5] facebook/wav2vec2-large-xlsr-53
[6] EdoAbati/whisper-small-it
[7] jonatasgrosman/exp_w2v2t_it_wav2vec2_s692
[8] jonatasgrosman/wav2vec2-large-xlsr-53-italian
[9] https://commonvoice.mozilla.org/it/datasets

Wav2Vec 2.0 Large (A) (68.50) achieves the best performance, followed by Wav2Vec 2.0 Base (A) (67.61), Whisper Small (A) (65.07), WavLM Base+ (A) (64.92), and XLSR-53 (A) (40.44). Examining the individual datasets in detail, we notice that the results on EMOVO are considerably lower than on the other two datasets, likely due to the small volume of training data. Whisper Small (A) outperforms the other models on DEMoS (93.09), while Wav2Vec 2.0 Large (A) and Wav2Vec 2.0 Base (A) yield the best results on EMOVO (48.84) and Emozionalmente (66.26), respectively. Interestingly, all English-only backbones perform comparably to or better than the multilingual ones (Whisper Small and XLSR-53). This indicates that English-only backbones are effective at extracting features useful for Italian SER, even if they have not been trained to handle cross-lingual differences. We hypothesize that this may be because SER relies on paralinguistic features, which are likely shared between English and Italian. Further investigation is needed regarding the competitive performance of English-only speech models compared to multilingual ones in Italian SER.

Regarding scenario (B), we investigate the impact of further fine-tuning the backbones with a self-supervised approach before addressing the Italian SER task. We find that this approach has a positive effect on XLSR-53, improving its F1 score from 40.44 (A) to 46.87 (B). However, it has a negative impact on both Whisper Small, with its F1 score decreasing from 65.07 (A) to 62.41 (B), and Wav2Vec 2.0 Large, with its F1 score dropping from 68.50 to 34.40. The significant drop for Wav2Vec 2.0 Large may be due to the substantial data shift that occurs when fine-tuning an English-only backbone on Italian corpora, compared to fine-tuning a multilingual one. Although fine-tuning XLSR on Italian data enhances performance, it remains less effective than English-only models, which achieve the best results without needing to specialize in the Italian language. This suggests the potential for directly using English-only pre-trained speech models for Italian SER tasks without requiring intermediate fine-tuning steps.

Fine-tuning Scenario	Backbone	Downstream Model	+ Adapter	+ LoRA
A	**Wav2vec 2.0 Base**	67.61 (±1.35)	65.97 (±0.81)	**67.67** (±1.12)
	WavLM Base+	64.92 (±1.66)	65.32 (±1.87)	**67.45** (±1.71)
	Wav2vec 2.0 Large	68.50 (±0.45)	67.83 (±0.61)	**69.57** (±0.49)
	Whisper Small	65.07 (±1.80)	56.10 (±2.18)	**68.91** (±1.91)
	XLSR-53	40.44 (±1.97)	40.20 (±1.56)	**41.02** (±1.88)
B	**Wav2vec 2.0 Large**	**34.40** (±2.63)	14.42 (±3.56)	21.88 (±3.17)
	Whisper Small	**62.41** (±4.12)	43.45 (±5.76)	55.45 (±5.43)
	XLSR-53	**46.87** (±0.58)	42.64 (±1.76)	43.94 (±0.65)

TABLE III

PERFORMANCE COMPARISONS BETWEEN DIFFERENT PEFT METHODS FOR SER. F1 SCORES ARE AVERAGED OVER THE THREE DATASETS (EMOVO, EMOZIONALMENTE, DEMOS) AND DIVIDED BY FINE-TUNING SCENARIOS (A) AND (B). VALUES ARE PRESENTED AS THE MEAN (STANDARD DEVIATION) ACROSS THREE RUNS.

2) PEFT Performance: To understand the impact of PEFT techniques, we compare them to the baseline downstream classification. In doing so, we average the F1 scores across the three datasets. Table III shows the mean F1 scores (± std) for each model across the three runs, distinguishing between scenario (A) and scenario (B). We observe that the Adapter approach provides slight improvement only for WavLM Base+ (A), while resulting in worse performance for all other models compared to direct downstream classification. This suggests that the Adapter is not an effective PEFT strategy compared to simply adding a downstream classification module on top of the speech backbone for Italian SER. In contrast, LoRA proves to be a more effective strategy for speech backbones in scenario (A) (e.g., Whisper Small (A) improves from 65.07 to 68.91). This indicates that adding task-specific parameters to the backbones using the LoRA method helps the backbone specialize in the task of Italian SER. For models in scenario (B), additional parameter tuning negatively impacts performance. This is evident with Whisper Small (B), where the F1 score drops from 62.41 to 55.45. We hypothesize that preliminary fine-tuning in Italian with an SSL approach may cause the model's features to lose general characteristics necessary for the paralinguistic task of Italian SER, making further fine-tuning on SER less effective. Consistent with the baselines, all speech models in scenario (A) except XLSR-53 outperform those in scenario (B) when LoRA is applied. Overall, English-only models perform comparably to or better than multilingual ones, with Wav2Vec 2.0 Large emerging as the top performer among them (69.57).

V. CONCLUSION

In this paper, we investigate the application of PEFT techniques on pre-trained speech models for the cross-lingual paralinguistic task of SER in the Italian language. We test Adapter and LoRA methods on multiple speech models and various datasets, finding that LoRA performs the best, with Wav2Vec 2.0 Large achieving the highest performance. Additionally, we find that English-only models are effective in transferring knowledge to Italian SER and perform comparably to or even better than multilingual models. The observed effectiveness of transferring knowledge from English-only models to the Italian SER task suggests commonalities in paralinguistic features between the two languages, highlighting the need for future investigations.

REFERENCES

[1] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211, 2004.

[2] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[3] Norhaslinda Kamaruddin and Abdul Wahab Abdul Rahman. Valence-arousal approach for speech emotion recognition system. In *2013 International Conference on Electronics, Computer and Computation (ICECCO)*, pages 184–187. IEEE, 2013.

[4] Klaus R Scherer. Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1-2):227–256, 2003.

[5] Donn Morrison, Ruili Wang, and Liyanage C De Silva. Ensemble methods for spoken emotion recognition in call-centres. *Speech communication*, 49(2):98–112, 2007.

[6] ZHAO Xiaoming, YANG Yijiao, and ZHANG Shiqing. Survey of deep learning based multimodal emotion recognition. *Journal of Frontiers of Computer Science & Technology*, 16(7):1479, 2022.

[7] Jianfeng Zhao, Xia Mao, and Lijiang Chen. Speech emotion recognition using deep 1d & 2d cnn lstm networks. *Biomedical signal processing and control*, 47:312–323, 2019.

[8] Tiantian Feng, Hanieh Hashemi, Murali Annavaram, and Shrikanth S Narayanan. Enhancing privacy through domain adaptive noise injection for speech emotion recognition. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 7702–7706. IEEE, 2022.

[9] 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.

[10] 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.

[11] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*, 2021.

[12] Seong-Gyun Leem, Daniel Fulford, Jukka-Pekka Onnela, David Gard, and Carlos Busso. Adapting a self-supervised speech representation for noisy speech emotion recognition by using contrastive teacher-student learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[13] Tiantian Feng, Rajat Hebbar, and Shrikanth Narayanan. Trust-ser: On the trustworthiness of fine-tuning pre-trained speech embeddings for speech emotion recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11201–11205. IEEE, 2024.

[14] Daria Diatlova, Anton Udalov, Vitalii Shutov, and Egor Spirin. Adapting wavlm for speech emotion recognition. *arXiv preprint arXiv:2405.04485*, 2024.

[15] Tiantian Feng and Shrikanth Narayanan. Peft-ser: On the use of parameter efficient transfer learning approaches for speech emotion recognition using pre-trained speech models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2023.

[16] Yingting Li, Ambuj Mehrish, Rishabh Bhardwaj, Navonil Majumder, Bo Cheng, Shuai Zhao, Amir Zadeh, Rada Mihalcea, and Soujanya Poria. Evaluating parameter-efficient transfer learning approaches on sure benchmark for speech understanding. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[17] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[18] 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.

[19] 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.

[20] Irene Mantegazza and Stavros Ntalampiras. Italian speech emotion recognition. In *2023 24th International Conference on Digital Signal Processing (DSP)*, pages 1–5. IEEE, 2023.

[21] Fabio Catania. Speech emotion recognition in italian using wav2vec 2. *Authorea Preprints*, 2023.

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[23] Hicham Atassi, Maria Teresa Riviello, Zdeněk Smékal, Amir Hussain, and Anna Esposito. Emotional vocal expressions recognition using the cost 2102 italian database of emotional speech. *Development of Multimodal Interfaces: Active Listening and Synchrony: Second COST 2102 International Training School, Dublin, Ireland, March 23-27, 2009, Revised Selected Papers*, pages 255–267, 2010.

[24] J Ancilin and A Milton. Improved speech emotion recognition with mel frequency magnitude coefficient. *Applied Acoustics*, 179:108046, 2021.

[25] Arianna Mencattini, Eugenio Martinelli, Giovanni Costantini, Massimiliano Todisco, Barbara Basile, Marco Bozzali, and Corrado Di Natale. Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure. *Knowledge-Based Systems*, 63:68–81, 2014.

[26] Youddha Beer Singh and Shivani Goel. A lightweight 2d cnn based approach for speaker-independent emotion recognition from speech with new indian emotional speech corpora. *Multimedia Tools and Applications*, 82(15):23055–23073, 2023.

[27] Alexander Wurst, Michael Hopwood, Sifan Wu, Fei Li, and Yu-Dong Yao. Deep learning for the detection of emotion in human speech: The impact of audio sample duration and english versus italian languages. In *2023 32nd Wireless and Optical Communications Conference (WOCC)*, pages 1–6. IEEE, 2023.

[28] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Emotion recognition in speech using cross-modal transfer in the wild. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 292–301, 2018.

[29] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020.

[30] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[31] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[32] 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. IEEE, 2015.

[33] Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*, 2021.

[34] 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. *arXiv preprint arXiv:2101.00390*, 2021.

[35] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

[36] 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.

[37] Zeyu Han, Chao Gao, Jinyang Liu, Sai Qian Zhang, et al. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.

[38] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021.

[39] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[40] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024.

[41] Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, Massimiliano Todisco, et al. Emovo corpus: an italian emotional speech database. In *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*, pages 3501–3504. European Language Resources Association (ELRA), 2014.

[42] Emilia Parada-Cabaleiro, Giovanni Costantini, Anton Batliner, Maximilian Schmitt, and Björn W Schuller. Demos: An italian emotional speech corpus: Elicitation methods, machine learning, and perception. *Language Resources and Evaluation*, 54(2):341–383, 2020.