Using Large Speech Models for Feature Extraction in Cross-Lingual Speech Emotion Recognition

Federico D'Asaro[1,2,*,†], Juan José Márquez Villacís[1,†], Giuseppe Rizzo[1,2] and Andrea Bottino[2]

[1] *LINKS Foundation – AI, Data & Space (ADS)*

[2] *Politecnico di Torino – Dipartimento di Automatica e Informatica (DAUIN)*

Abstract

Large Speech Models (LSMs), pre-trained on extensive unlabeled data using Self-Supervised Learning (SSL) or Weakly-Supervised Learning (WSL), are increasingly employed for tasks like Speech Emotion Recognition (SER). Their capability to extract general-purpose features makes them a strong alternative to low-level descriptors. Most studies focus on English, with limited research on other languages. We evaluate English-Only and Multilingual LSMs from the Wav2Vec 2.0 and Whisper families as feature extractors for SER in eight languages. We have stacked three alternative downstream classifiers of increasing complexity, named *Linear, Non-Linear, and Multi-Layer*, on top of the LSMs. Results indicate that Whisper models perform best with a simple linear classifier using features from the last transformer layer, while Wav2Vec 2.0 models benefit from features from the middle and early transformer layers. When comparing English-Only and Multilingual LSMs, we find that Whisper models benefit from multilingual pre-training, excelling in Italian, Canadian French, French, Spanish, German and competitively on Greek, Egyptian Arabic, Persian. In contrast, English-Only Wav2Vec 2.0 models outperform their multilingual counterpart, XLS-R, in most languages, achieving the highest performance in Greek, Egyptian Arabic.

Keywords

Cross-lingual Speech Emotion Recognition, Large Speech models, Transfer Learning

1. Introduction

Speech Emotion Recognition (SER) aims to identify emotions from speech audio, enhancing Human-AI interaction in fields such as healthcare, education, and security [1]. Traditional methods rely on Low-Level Descriptors (LLD) like spectral, prosodic, and voice quality features [2], using classifiers such as KNN, SVM, or Naïve Bayes [3]. Deep learning has introduced advanced techniques, including Convolutional Neural Networks (CNNs) [4, 5, 6], eventually followed by Recurrent Neural Networks (RNNs) [7, 8], and Transformers [9, 10, 11]. Transformers' ability to learn from extensive datasets has led to Large Speech Models (LSMs), which generalize across various speech tasks. Common training approaches for these models include Self-Supervised Learning (SSL), which uses data itself to learn general-purpose features [12], and Weakly-Supervised Learning (WSL), which pairs audio with text for tasks like transcription and translation [13]. The general-purpose knowledge of LSMs makes them effective feature extractors for SER. Research has adapted LSMs for SER in English [14, 15, 16, 17], but efforts for other languages are limited, focusing on Wav2Vec 2.0 [18] for cross-lingual SER [19, 20, 21].

This study examines how effective LSMs are as feature extractors for cross-lingual SER, using nine datasets across eight languages: *Italian, German, French, Canadian French, Spanish, Greek, Persian, and Egyptian Arabic.* Specifically, we utilize LSMs from the Wav2Vec 2.0 and Whisper [13] model families, pre-trained with SSL and WSL approaches, respectively. We introduce Whisper due to its underexplored use in cross-lingual SER. To assess the effectiveness of LSMs as feature extractors, we test three classifiers of increasing complexity—*Linear, Non-Linear, and Multi-Layer*—across nine datasets. This evaluation determines which classifier best suits each LSM across different languages. Moreover, our study includes both English-Only and Multilingual models from the Wav2Vec 2.0 and Whisper families, aiming to evaluate the effectiveness of multilingual pre-training for cross-lingual SER.

The main contributions of this work are:

- We evaluate LSMs from the Wav2Vec 2.0 and Whisper models as feature extractors for cross-lingual SER across eight languages.
- We test three types of downstream classifiers—Linear, Non-Linear, and Multi-Layer—and find that Whisper models' last Transformer layer features are well-suited for a Linear classifier, whereas Wav2Vec 2.0 models perform better with

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

[*] Corresponding author.

[†] These authors contributed equally.

federico.dasaro@polito.it (F. D'Asaro); juan.marquez@linksfoundation.com (J. J. Márquez Villacís); giuseppe.rizzo@linksfoundation.com (G. Rizzo); andrea.bottino@polito.it (A. Bottino)

http://conceptbase.sourceforge.net/mjf/ (A. Bottino)

0009-0003-8727-3393 (F. D'Asaro); 0009-0008-3098-5492 (J. J. Márquez Villacís); 0000-0003-0083-813X (G. Rizzo); 0000-0002-8894-5089 (A. Bottino)

features from the middle and early Transformer layers.

- We compare English-Only and Multilingual LSMs, revealing that Whisper models benefit from multilingual pre-training performing best on Italian, Spanish, Canadian French, French, and German and competitively on Greek, Egyptian Arabic, Persian. Conversely, English-Only Wav2Vec 2.0 models surpass multilingual XLS-R in most languages, achieving the highest performance in Greek, Egyptian Arabic.

2. Background

2.1. Large Speech Models

Recent developments in natural language processing and computer vision have harnessed large volumes of unlabeled data through Self-Supervised Learning [22, 23, 24]. Building on techniques such as masked language and image modeling, Wav2Vec 2.0 [18] introduced a LSM trained on extensive audio datasets using masked speech modeling. Wav2Vec 2.0 features seven 1D convolutional blocks for initial feature extraction, followed by 12 or 24 transformer blocks (depending on the model variant) for contextual processing. The model masks part of the latent features and reconstructs them using the surrounding context. To further refine LSMs for tasks like emotion recognition, methods such as WavLM [25] have been developed. WavLM incorporates speech denoising alongside masked modeling, demonstrating broad effectiveness across various tasks in the SUPERB benchmark [26]. Moreover, XLSR-53 [27] extends the Wav2Vec 2.0 framework to cover 53 languages, sharing the latent space across these languages. This approach has shown superior performance over monolingual pretraining for automatic speech recognition. XLS-R [28] further advances this by scaling to 128 languages, excelling in speech translation and language identification. In comparison, Whisper [13] leverages large-scale weak supervision from audio-transcription pairs to train an encoder-decoder transformer. Using log-mel spectrograms, Whisper is trained in a multitask framework that includes multilingual transcription and translation, establishing itself as an effective zero-shot model for multilingual tasks.

2.2. Cross-Language Speech Emotion Recognition

Emotion recognition in languages beyond English, like Italian [29], French [30], Persian [31, 32], and Spanish [33], is crucial but often limited by data availability. Recent efforts have focused on improving cross-lingual and cross-modal knowledge transfer. Techniques like dual attention [21] and tensor fusion [34] enhance audio and text interaction in languages such as Italian, German, and Urdu. Self-supervised pre-training methods, including variational autoencoders, have also been effective in transferring knowledge across languages like German [35, 36]. The advent of LSMs pre-trained with self-supervision has further increased the potential for transfer learning due to their high generalization capabilities [15]. However, most research primarily focuses on adapting multilingual Wav2Vec 2.0 models (XLSR-53) [19, 37, 20, 21]. This work expands the scope of analyzed LSMs including WSL models as Whisper. Additionally, we evaluate the ability of English-only models to transfer knowledge to other languages, beyond just multilingual models.

3. Method

In this section, we describe the methodology for evaluating the effectiveness of LSMs as feature extractors for downstream SER in various languages. We stack a classification model on top of the LSM backbone, with its parameters frozen. All LSMs used in this work share the same overall architecture, which we describe below along with the stacked classification model.

Formally, the input audio A (raw waveform or log-mel spectrogram) passes through a convolutional encoder $\mathcal{z} : A \rightarrow Z$, mapping the audio to latent features $Z = \{z_1, \dots, z_T\}$, where T is the sequence length and each frame z_i typically corresponds to 25 ms with $z_i \in \mathbb{R}^d$. Then, Z passes through a Transformer encoder consisting of l layers $\mathcal{h}^l : Z \rightarrow H$, enriching the latent features with contextual information, resulting in $\{h_1^l, \dots, h_T^l\}$ for each of the $l = 1, \dots, L$ Transformer layers. Here, $l = L$ corresponds to the output features of the last layer, with $h_i^l \in \mathbb{R}^d$. The features $\{h_1^l, \dots, h_T^l\}_{l=1,..,L}$ are considered the extracted features from the LSM and are fed into a downstream classifier $\mathcal{y} : H \rightarrow Y$, which maps these features to the output class logits $\{y_1, \dots, y_k\}$. The output class label y^* for audio A is given by:

$$y^* = \arg\max_k \operatorname{softmax}\left(\mathcal{y}\left(\mathcal{h}\left(\mathcal{z}(A)\right)\right)\right) \qquad (1)$$

Inspired by previous work that uses probing to evaluate the quality of features extracted from backbone models [38, 39], we evaluate three different downstream classifiers of increasing complexity: *Linear Classifier* ($\mathcal{g}_l$), *Non-Linear Classifier* ($\mathcal{g}_{nl}$), and *Multi-layer Classifier* ($\mathcal{g}_{ml}$). Figure 1 illustrates their architecture, which is detailed below.

3.1. Linear Classifier

For the linear classifier, we use a simple feed-forward neural network that consists solely of linear projections.

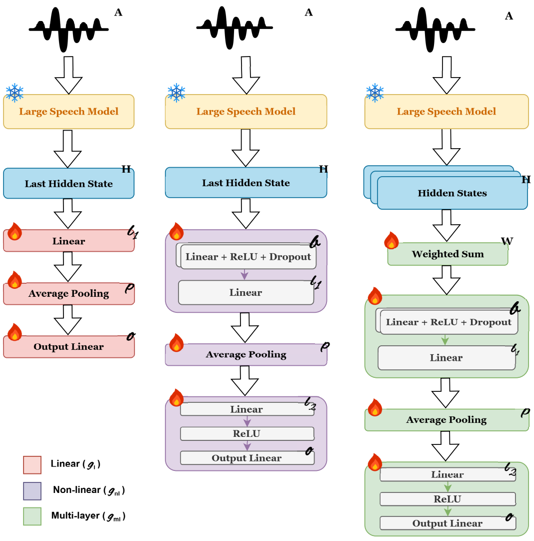

Figure 1: The three downstream classifiers used in this work are: Linear (red), Non-Linear (purple), and Multi-Layer (green). The snowflake icon represents frozen weights, while the fire icon denotes trainable weights.

Specifically, given the features from the last Transformer layer $\{h_1^L, \ldots, h_T^L\}$, they are first projected by a linear layer $\ell_1 : \mathbb{R}^d \rightarrow \mathbb{R}^m$ that is shared across all frames, then aggregated by average pooling $\mathcal{p}$, and finally pass through the classification layer $\mathcal{o} : \mathbb{R}^m \rightarrow \mathbb{R}^k$ to obtain the output class logits. The function $\mathcal{g}_l$ is compactly defined as:

$$\mathcal{g}_l\left(h_1^L, \ldots, h_T^L\right) = \mathcal{o}\left(\mathcal{p}\left(\ell_1\left(h_1^L, \ldots, h_T^L\right)\right)\right) \tag{2}$$

The absence of non-linear activations allows us to evaluate the quality of the features extracted from the LSM based on the linear classifier model's ability to handle the SER task.

3.2. Non-Linear Classifier

To increase the complexity of the classification model, we utilize a series of linear layers interleaved with ReLU activations both before and after feature pooling. We follow the same architecture as in [14, 15], but unlike them, we only feed the features from the last Transformer layer L to the model. Each $\{h_1^L, \ldots, h_T^L\}$ passes through two shared linear layers, ReLU, and dropout blocks ($\mathcal{b}$), followed by a linear layer (ℓ_1). Linear layers are functions $\ell : \mathbb{R}^d \rightarrow \mathbb{R}^m$. Projected features are averaged, pass through ℓ_2 and ReLU, and are classified by $\mathcal{o}$. Thus, $\mathcal{g}_{nl}$ is:

$$\mathcal{g}_{nl}\left(x = h_1^L, \ldots, h_T^L\right) = \mathcal{o}\left(\mathrm{ReLU}\left(\ell_2\left(\mathcal{p}\left(\ell_1\left(\mathcal{b}\left(x\right)\right)\right)\right)\right) \tag{3}$$

3.3. Multi-Layer Classifier

As a third option, we adopt the approach from [14, 15], which utilizes all hidden states of the Transformer encoder. The features $\{h_1^l, \ldots, h_T^l\}_{l=1,\ldots,L}$ are combined into a new sequence $\{h_1^*, \ldots, h_T^*\}$ using a learnable weighted sum. The function $\mathcal{s} : \mathbb{R}^{L \times T \times d} \rightarrow \mathbb{R}^{T \times d}$ maps $\{h_1^l, \ldots, h_T^l\}_{l=1,\ldots,L}$ to $\{h_1^*, \ldots, h_T^*\}$ as follows:

$$h_t^* = \sum_{l=1}^{L} w_l \cdot h_t^l \quad \text{for } t = 1, \ldots, T \tag{4}$$

where $w_1, \ldots, w_L$ are the weights assigned to each Transformer layer, ensuring $w_l \in [0, 1]$ and $\sum_{l=1}^{L} w_l = 1$. The resulting sequence $\{h_1^*, \ldots, h_T^*\}$ is then processed by the same pipeline as the *Non-Linear Classifier*, resulting in:

$$\mathcal{g}_{ml}\left(x = \{h_1^l, \ldots, h_T^l\}_{l=1,\ldots,L}\right) = \mathcal{g}_{nl}\left(\mathcal{s}(x)\right) \tag{5}$$

This classifier leverages internal layer information, which has proven beneficial for paralinguistic and linguistic downstream tasks [39, 40, 41, 42]. By investigating the contribution of internal LSM layers for SER across various languages, we corroborates previous findings for Wav2Vec 2.0 models and provide new insights for Whisper models.

4. Experiments

4.1. Datasets and Metrics

In this study, we conduct experiments using 9 distinct datasets spanning 8 different languages: *Greek, French, Italian, German, Spanish, Egyptian Arabic, and Persian.* The datasets vary in their collection methodologies, such as acted emotions and elicitation methods. The participant demographics may be balanced by gender (e.g., CaFE, EYASE), by emotion (e.g., EMOVO), or may not be balanced at all. For all datasets, we conduct our experiments in a speaker-independent setting to prevent evaluation on speaker-dependent features. Table 1 provides an overview of the dataset statistics, with a more detailed description given below.

AESDD [43]: The Acted Emotional Speech Dynamic Database comprises 500 recorded samples from 5 actors (3 females, 2 males) expressing 5 distinct emotions in Greek. Each actor performed 20 utterances per emotion, with some utterances recorded multiple times. In later versions, additional actors were included, bringing the total to 604 recordings from 6 actors.

CaFE [44]: This dataset includes recordings of 6 different sentences delivered by 12 actors (6 female, 6 male) portraying the Big Six emotions and a neutral state in Canadian French. It offers a high-quality version with a sampling rate of 192 kHz at 24 bits per sample, as well as

Dataset	Language	# Samples	Emotions
AESDD	Greek	500	anger, disgust, fear, happiness, and sadness
CaFE	Canadian French	936	anger, disgust, fear, happiness, surprise, sadness, and neutrality
DEMoS	Italian	9697	anger, disgust, fear, happiness, surprise, sadness, and neutrality
EmoDB	German	535	anger, disgust, fear, happiness, boredom, sadness, and neutrality
EmoMatch	Spanish	2005	anger, disgust, fear, happiness, surprise, sadness, and neutrality
EMOVO	Italian	588	anger, disgust, fear, happiness, surprise, sadness, and neutrality
EYASE	Egyptian Arabic	579	anger, happiness, sadness, and neutrality
Oréau	French	502	anger, disgust, fear, happiness, surprise, sadness, and neutrality
ShEMO	Persian	400	anger, happiness, sadness, and neutrality

Table 1
Summary statistics of the 9 datasets used in this work.

a down-sampled version at 48 kHz and 16 bits per sample. The total number of samples amounts to 936.

DEMoS [45]: DEMoS contains 9697 audio samples from 68 volunteer students (299 females, 131 males) expressing the Big Six emotions plus the neutral state in Italian. Instead of acted emotions, samples were generated using an elicitation approach. The recordings, with a mean duration of 2.9 seconds (std: 1.1s), are provided in 48 kHz, 16-bit, mono format.

EmoDB [46]: This collection includes 535 utterances across 7 emotional states, spoken in German by 5 female and 5 male actors. Each actor performed a set of 10 sentences, which were down-sampled from the original 48 kHz to 16 kHz.

EmoMatch [33]: Consisting of 2005 recordings, EmoMatch features samples from 50 non-actor Spanish speakers (20 females, 30 males) expressing the Big Six emotions and a neutral state. The dataset is a subset of the larger EmoSpanishDB and contains recordings sampled at 48 kHz with a 16-bit mono format.

EMOVO [47]: EMOVO presents 588 Italian audio recordings from 3 male and 3 female actors simulating the Big Six emotions plus a neutral state. Each actor voiced 14 utterances, and the recordings are provided in 48 kHz, 16-bit stereo WAV format.

EYASE [48]: EYASE contains 579 utterances in Egyptian Arabic, recorded by 3 male and 3 female professional actors. The recordings, ranging from 1 to 6 seconds in duration, were labeled as angry, happy, neutral, or sad and sampled at 44.1 kHz.

Oréau [49]: The Oréau dataset features 502 audio samples from 32 non-professional actors (25 male, 7 female) who voiced 10 to 13 utterances in French for the Big Six emotions plus a neutral state.

ShEMO [50]: ShEMO comprises 3000 semi-natural recordings from 87 native Persian speakers (31 female, 56 male). The dataset captures 5 of the Big Six emotions—sadness, anger, happiness, surprise, and fear—plus a neutral state. The samples were up-sampled to a frequency of 44.1 kHz in mono-channel format, with an average length of 4.11 seconds (std: 3.41s).

The audio is resampled to 16 kHz, and a stratified train/-validation/test split is performed with ratios of 80/10/10. All results are reported using the macro F1 score, expressed as a percentage. We conducted 3 runs, presenting the mean ± standard deviation.

4.2. Experimental Details

Baseline As a baseline to evaluate LSM transfer learning capabilities, we adopt the Audio Spectrogram Transformer (AST) [51], a fully transformer-based architecture recently proposed as a substitute for CNNs [9, 10, 11]. We train AST from scratch on each of the 9 datasets using the same hyperparameters as [51].

LSM Models We use pre-trained checkpoints for both English-Only and Multilingual models: Wav2Vec 2.0 Base, Wav2Vec 2.0 Large, XLS-R from the Wav2Vec 2.0 family, and Whisper Small (EN) (Whisper Small pre-trained only on English data), Whisper Small, Whisper Medium from the Whisper family. The LSM backbones are kept frozen and used exclusively as feature extractors.

Training We follow the same hyperparameters settings as [15] to train the downstream classifiers. Specifically, we train for 30 epochs using the Adam optimizer with a learning rate of 5.0e-04, weight decay of 1.0e-04, betas set to (0.9, 0.98), and epsilon of 1.0e-08. The dimension of the classifier projection m is 256.

4.3. Results

To present our results, we first compare the performance of the various classifiers (see Section 3) for each LSM utilized. This analysis provides insights into the characteristics of features extracted from Wav2Vec 2.0 and Whisper models for downstream SER tasks. After identifying the best classifier for each LSM, we then compare the performance of English-Only and Multilingual LSMs across the 8 languages covered in this study.

4.3.1. Comparison between downstream classifiers

We examine the results in Table 2, comparing three classifier methods for Wav2Vec 2.0 and Whisper models. The

Backbone	Linear	Non-Linear	Multi-Layer
Wav2Vec 2.0 Base	47.87 (± 0.93)	42.07 (± 5.27)	**53.42** (± 1.27)
Wav2Vec 2.0 Large	12.09 (± 1.50)	12.93 (± 3.31)	**57.50** (± 0.03)
XLS-R	5.43 (± 0.40)	5.86 (± 0.07)	**40.89** (± 2.00)
Whisper Small (EN)	**58.16** (± 0.15)	53.50 (± 0.98)	49.73 (± 2.02)
Whisper Small	**60.87** (± 0.26)	54.86 (± 0.93)	45.14 (± 1.54)
Whisper Medium	**60.72** (± 0.16)	55.56 (± 1.09)	37.95 (± 2.27)

Table 2
Performance of various LSM backbones using *Linear*, *Non-Linear*, and *Multi-Layer* classification methods. F1 scores are averaged across all 9 datasets. For each LSM, the best classifier is highlighted in bold.

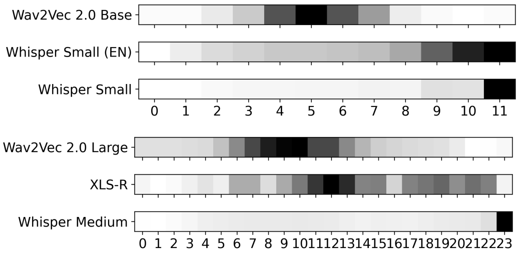

Figure 2: Greyscale map of layer weight distribution from the *Multi-Layer* classification method. Weights are averaged over all 9 datasets for each model. Darker shades indicate higher weights.

table shows average F1 scores across 9 datasets, highlighting the most effective classifier for each LSM in cross-lingual SER tasks.

For Wav2Vec 2.0 models, the Multi-Layer Classifier performs best, with F1 scores of 53.42, 57.50, and 40.89 for Wav2Vec 2.0 Base, Wav2Vec 2.0 Large, and XLS-R. The Linear and Non-Linear classifiers perform similarly, especially for Wav2Vec 2.0 Large and XLS-R, suggesting improvements are due to using features from internal Transformer layers rather than non-linear activations. For Whisper models, the Linear Classifier performs best, with F1 scores of 58.16, 60.87, and 60.72 for Whisper Small (EN), Whisper Small, and Whisper Medium. Increasing classifier complexity with non-linear activations decreases performance, likely due to general information loss caused by complex transformations. The Multi-Layer Classifier performs worse, indicating that using also features from internal layers is less effective than using features from the last layer alone.

This comparison reveals that Wav2Vec 2.0 models benefit from features extracted from internal Transformer layers and exhibit less sensitivity to classifier complexity, consistent with prior research [41, 39]. Conversely, Whisper models achieve better performance with features from the last Transformer layer when using a simple linear classifier, offering new insights into their effectiveness for SER across multiple languages. We hypothesize that this differing behavior may be related to their respective Self-Supervised and Weakly-Supervised pre-training approaches, which warrant further investigation. To gain further insights into the importance of Transformer layers in Wav2Vec 2.0 and Whisper for SER, we leverage the weights learned in the Multi-Layer classifier as follows.

Transformer Layer Weights. We analyze the weights $w_1, \ldots, w_L$ from the Multi-Layer Classifier to assess Transformer layer importance. Figure 2 illustrates that Wav2Vec 2.0 models assign greater weight to the early and middle layers, whereas Whisper models emphasize the later layers. This observation confirms the earlier findings, suggesting that paralinguistic information in Whisper models is embedded in the features of the later Transformer layers.

4.3.2. Comparing English-Only and Multilingual LSMs Across Different Languages

In this section, we compare English-Only and Multilingual LSMs with the AST baseline across 9 datasets. Table 3 displays F1 scores for the optimal classifiers found in the previous section: *Multi-Layer* for Wav2Vec 2.0 and *Linear* for Whisper models.

Transferring knowledge from LSMs proves to be effective across all datasets compared to the baseline. For instance, Wav2Vec 2.0 Large scores 53.40 in Egyptian Arabic, while Whisper Small scores 51.98 and AST scores 33.23. This indicates that LSMs are effective feature extractors for cross-lingual SER on multiple languages.

When comparing English-only and Multilingual models, we differentiate between the Wav2Vec 2.0 and Whisper families. For Wav2Vec 2.0, we observe that Wav2Vec 2.0 Base and Large generally outperform XLS-R (e.g., 87.85 and 88.31 vs. 67.71 for DEMos), except in Persian, where their performance is comparable. This indicates that multilingual pre-training may not be as effective for Wav2Vec 2.0 models across various languages. We speculate that this may be due to the limitations of SSL pre-training, which might struggle with the diverse range of languages and lose important paralinguistic features that are retained in English-only models. Further investigation with a wider range of SSL-pretrained LSMs could provide more insights. As regards to Whisper, Multilingual Whisper Small outperforms its English-only version, with the exception of Greek and Persian, likely due to limited pretraining data for these languages, which resulted in higher word error rates compared to other languages in this study [13]. Multilingual Whisper models achieve best performance in Canadian French, Spanish (66.71, 73.13 with Whisper Small), Italian, German, and French (91.17, 90.64, 95.22 with Whisper Medium). This improvement is likely due to the larger pretraining datasets for these languages and the similarities between

		English-Only			Multilingual		
Dataset/Model	AST	Wav2Vec 2.0 Base‡	Wav2Vec 2.0 Large‡	Whisper Small†	XLS-R‡	Whisper Small†	Whisper Medium†
AESDD (el)	19.84 (± 0.16)	25.45 (± 0.98)	**28.89** (± 2.64)	28.04 (± 0.99)	9.16 (± 1.25)	26.34 (± 1.65)	27.62 (± 0.62)
CaFE (fr-ca)	10.96 (± 6.26)	50.52 (± 3.54)	47.74 (± 0.33)	60.66 (± 0.76)	18.66 (± 0.01)	**66.71** (± 0.72)	55.03 (± 0.38)
DEMoS (it)	13.75 (± 4.26)	87.85 (± 0.01)	88.31 (± 0.74)	88.24 (± 0.21)	67.71 (± 1.47)	90.61 (± 0.14)	**91.17** (± 0.20)
EmoDB (de)	46.11 (± 6.55)	81.75 (± 7.30)	88.84 (± 7.48)	83.31 (± 0.18)	67.39 (± 4.33)	87.21 (± 1.11)	**90.64** (± 1.47)
EmoMatch (es)	36.10 (± 2.63)	69.84 (± 0.69)	71.85 (± 1.55)	67.59 (± 0.35)	44.14 (± 0.25)	**73.13** (± 2.54)	68.23 (± 0.78)
EMOVO (it)	15.74 (± 1.24)	16.47 (± 0.61)	20.33 (± 1.31)	27.30 (± 0.16)	14.86 (± 2.11)	41.05 (± 1.21)	**50.19** (± 0.29)
EYASE (ar-eg)	33.23 (± 4.58)	46.31 (± 3.62)	**53.40** (± 1.56)	42.65 (± 0.70)	47.27 (± 1.36)	51.98 (± 0.88)	37.32 (± 3.62)
Oréau (fr)	19.01 (± 2.35)	52.86 (± 0.07)	58.42 (± 4.14)	82.27 (± 0.23)	32.51 (± 4.89)	92.70 (± 1.67)	**95.22** (± 0.84)
ShEMO (fa)	36.15 (± 0.85)	60.55 (± 3.90)	57.52 (± 9.09)	**67.93** (± 0.37)	61.24 (± 8.93)	63.88 (± 1.21)	63.85 (± 1.58)

Table 3

Performance of Wav2Vec and Whisper models across 9 datasets, divided into English-Only and Multilingual LSMs. AST is the baseline. † indicates a Linear Classifier, ‡ a Multi-Layer Classifier. Bold values are the highest scores, and underlined values highlight the best between English-Only and Multilingual models.

Canadian French and French. We believe that multilingual pretraining benefits Whisper models by capturing language-specific features more effectively through WSL and multitask learning. However, further research is needed to evaluate the effectiveness of multilingual pretraining with WSL compared to SSL across a broader range of LSMs.

5. Conclusion

This paper examines the capabilities of Wav2Vec 2.0 and Whisper models as feature extractors for cross-lingual SER across eight languages, considering both English-Only and Multilingual variants. Our findings reveal that LSMs are effective feature extractors compared to a full Transformer baseline trained from scratch. We observe that Whisper models encode acoustic information primarily in the features of the last Transformer layer, whereas Wav2Vec 2.0 models rely on features from middle and early layers. Furthermore, we show that multilingual pre-training benefits Whisper models, leading to strong performance in Italian, Canadian French, French, Spanish, German, and competitive results in Greek, Egyptian Arabic, and Persian. In contrast, English-Only Wav2Vec 2.0 models outperform their multilingual counterpart, XLS-R, in most languages, achieving top performance in Greek and Egyptian Arabic. We attribute the disparity in multilingual pre-training effectiveness to the differences between SSL and WSL strategies, which should be explored further.

References

[1] C.-C. Lee, K. Sridhar, J.-L. Li, W.-C. Lin, B.-H. Su, C. Busso, Deep representation learning for affective speech signal analysis and processing: Preventing unwanted signal disparities, IEEE Signal Processing Magazine 38 (2021) 22–38.

[2] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, Y. Zong, A survey of deep learning-based multimodal emotion recognition: Speech, text, and face, Entropy 25 (2023) 1440.

[3] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, E. Ambikairajah, A comprehensive review of speech emotion recognition systems, IEEE access 9 (2021) 47795–47814.

[4] Z. Huang, M. Dong, Q. Mao, Y. Zhan, Speech emotion recognition using cnn, in: Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 801–804.

[5] A. M. Badshah, J. Ahmad, N. Rahim, S. W. Baik, Speech emotion recognition from spectrograms with deep convolutional neural network, in: 2017 international conference on platform technology and service (PlatCon), IEEE, 2017, pp. 1–5.

[6] J. Zhao, X. Mao, L. Chen, Speech emotion recognition using deep 1d & 2d cnn lstm networks, Biomedical signal processing and control 47 (2019) 312–323.

[7] T. Feng, H. Hashemi, M. Annavaram, S. 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), IEEE, 2022, pp. 7702–7706.

[8] W. Lim, D. Jang, T. Lee, Speech emotion recognition using convolutional and recurrent neural networks,

in: 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA), IEEE, 2016, pp. 1–4.

[9] N.-C. Ristea, R. T. Ionescu, F. S. Khan, Septr: Separable transformer for audio spectrogram processing, arXiv preprint arXiv:2203.09581 (2022).

[10] J.-Y. Kim, S.-H. Lee, Coordvit: a novel method of improve vision transformer-based speech emotion recognition using coordinate information concatenate, in: 2023 International conference on electronics, information, and communication (ICEIC), IEEE, 2023, pp. 1–4.

[11] S. Akinpelu, S. Viriri, A. Adegun, An enhanced speech emotion recognition using vision transformer, Scientific Reports 14 (2024) 13126.

[12] S. Liu, A. Mallol-Ragolta, E. Parada-Cabaleiro, K. Qian, X. Jing, A. Kathan, B. Hu, B. W. Schuller, Audio self-supervised learning: A survey, Patterns 3 (2022).

[13] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: International Conference on Machine Learning, PMLR, 2023, pp. 28492–28518.

[14] L. Pepino, P. Riera, L. Ferrer, Emotion recognition from speech using wav2vec 2.0 embeddings, arXiv preprint arXiv:2104.03502 (2021).

[15] T. Feng, S. 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), IEEE, 2023, pp. 1–8.

[16] Y. Li, A. Mehrish, R. Bhardwaj, N. Majumder, B. Cheng, S. Zhao, A. Zadeh, R. Mihalcea, S. 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), IEEE, 2023, pp. 1–5.

[17] T. Feng, R. Hebbar, S. 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), IEEE, 2024, pp. 11201–11205.

[18] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems 33 (2020) 12449–12460.

[19] M. Sharma, Multi-lingual multi-task speech emotion recognition using wav2vec 2.0, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 6907–6911.

[20] S. G. Upadhyay, L. Martinez-Lucas, B.-H. Su, W.-C. Lin, W.-S. Chien, Y.-T. Wu, W. Katz, C. Busso, C.-C. Lee, Phonetic anchor-based transfer learning to facilitate unsupervised cross-lingual speech emotion recognition, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[21] S. A. M. Zaidi, S. Latif, J. Qadir, Cross-language speech emotion recognition using multimodal dual attention transformers, arXiv preprint arXiv:2306.13804 (2023).

[22] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[23] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018).

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[25] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al., Wavlm: Large-scale self-supervised pre-training for full stack speech processing, IEEE Journal of Selected Topics in Signal Processing 16 (2022) 1505–1518.

[26] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, et al., Superb: Speech processing universal performance benchmark, arXiv preprint arXiv:2105.01051 (2021).

[27] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, M. Auli, Unsupervised cross-lingual representation learning for speech recognition, arXiv preprint arXiv:2006.13979 (2020).

[28] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. Von Platen, Y. Saraf, J. Pino, et al., Xls-r: Self-supervised cross-lingual speech representation learning at scale, arXiv preprint arXiv:2111.09296 (2021).

[29] A. Wurst, M. Hopwood, S. Wu, F. Li, Y.-D. 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), IEEE, 2023, pp. 1–6.

[30] M. Neumann, et al., Cross-lingual and multilingual speech emotion recognition on english and french, in: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2018, pp. 5769–5773.

[31] S. Deng, N. Zhang, Z. Sun, J. Chen, H. Chen, When

low resource nlp meets unsupervised language model: Meta-pretraining then meta-learning for few-shot text classification (student abstract), in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 13773–13774.

[32] S. Latif, A. Qayyum, M. Usman, J. Qadir, Cross lingual speech emotion recognition: Urdu vs. western languages, in: 2018 International conference on frontiers of information technology (FIT), IEEE, 2018, pp. 88–93.

[33] E. Garcia-Cuesta, A. B. Salvador, D. G. Pãez, Emomatchspanishdb: study of speech emotion recognition machine learning models in a new spanish elicited database, Multimedia Tools and Applications 83 (2024) 13093–13112.

[34] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, arXiv preprint arXiv:1707.07250 (2017).

[35] H. H. Mao, A survey on self-supervised pre-training for sequential transfer learning in neural networks, arXiv preprint arXiv:2007.00800 (2020).

[36] S. Sadok, S. Leglaive, R. Séguier, A vector quantized masked autoencoder for speech emotion recognition, in: 2023 IEEE International conference on acoustics, speech, and signal processing workshops (ICASSPW), IEEE, 2023, pp. 1–5.

[37] F. Catania, Speech emotion recognition in italian using wav2vec 2, Authorea Preprints (2023).

[38] Y. Belinkov, J. Glass, Analyzing hidden representations in end-to-end automatic speech recognition systems, Advances in Neural Information Processing Systems 30 (2017).

[39] J. Shah, Y. K. Singla, C. Chen, R. R. Shah, What all do audio transformer models hear? probing acoustic representations for language delivery and its structure, arXiv preprint arXiv:2101.00387 (2021).

[40] A. Pasad, J.-C. Chou, K. Livescu, Layer-wise analysis of a self-supervised speech representation model, in: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2021, pp. 914–921.

[41] Y. Li, Y. Mohamied, P. Bell, C. Lai, Exploration of a self-supervised speech model: A study on emotional corpora, in: 2022 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2023, pp. 868–875.

[42] A. Pasad, B. Shi, K. Livescu, Comparative layer-wise analysis of self-supervised speech models, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[43] N. Vryzas, R. Kotsakis, A. Liatsou, C. A. Dimoulas, G. Kalliris, Speech emotion recognition for performance interaction, Journal of the Audio Engineering Society 66 (2018) 457–467.

[44] P. Gournay, O. Lahaie, R. Lefebvre, A canadian french emotional speech dataset, in: Proceedings of the 9th ACM multimedia systems conference, 2018, pp. 399–402.

[45] E. Parada-Cabaleiro, G. Costantini, A. Batliner, M. Schmitt, B. W. Schuller, Demos: An italian emotional speech corpus: Elicitation methods, machine learning, and perception, Language Resources and Evaluation 54 (2020) 341–383.

[46] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, et al., A database of german emotional speech., in: Interspeech, volume 5, 2005, pp. 1517–1520.

[47] G. Costantini, I. Iaderola, A. Paoloni, M. Todisco, et al., Emovo corpus: an italian emotional speech database, in: Proceedings of the ninth international conference on language resources and evaluation (LREC'14), European Language Resources Association (ELRA), 2014, pp. 3501–3504.

[48] L. Abdel-Hamid, Egyptian arabic speech emotion recognition using prosodic, spectral and wavelet features, Speech Communication 122 (2020) 19–30.

[49] S. Oréau, French emotional speech database - oréau, Zenodo, 2021. URL: https://zenodo.org/records/4405783.

[50] O. Mohamad Nezami, P. Jamshid Lou, M. Karami, Shemo: a large-scale validated database for persian speech emotion detection, Language Resources and Evaluation 53 (2019) 1–16.

[51] Y. Gong, Y.-A. Chung, J. Glass, Ast: Audio spectrogram transformer, arXiv preprint arXiv:2104.01778 (2021).