2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS) | 979-8-3503-1890-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/SNAMS60348.2023.10375468

Breaking Barriers: Can Multilingual Foundation Models Bridge the Gap in Cross-Language Speech Emotion Recognition?

Moazzam Shoukat
EmulationAI
Pakistan
moazzam.shoukat@emulationai.com

Muhammad Usama
National University of Computer & Emerging Sciences, Faisalabad, Pakistan
m.usama@nu.edu.pk

Hafiz Shehbaz Ali
EmulationAI
Pakistan
shehbaz.ali@emulationai.com

Siddique Latif
University of Southern Queensland (UniSQ)
Queensland University of Technology (QUT)
Australia
0000-0001-5662-4777

***Abstract*—Speech emotion recognition (SER) faces challenges in cross-language scenarios due to differences in linguistic and cultural expression of emotions across languages. Recently, large multilingual foundation models pre-trained on massive corpora have achieved performance on natural language understanding tasks by learning cross-lingual representations. Their ability to understand relationships between languages without direct translation opens up possibilities for more applicable multilingual models. In this paper, we evaluate the capabilities of foundation models (Wav2Vec2, XLSR, Whisper and MMS) to bridge the gap in cross-language SER. Specifically, we analyse their performance on benchmark cross-language SER datasets involving four languages for emotion classification. Our experiments show that the foundation model outperforms CNN-LSTM baselines, establishing their superiority in cross-lingual transfer learning for emotion recognition. However, self-supervised pre-training plays a key role, and inductive biases alone are insufficient for high cross-lingual generalisability. Foundation models also demonstrate gains over baselines with limited target data and better performance on noisy data. Our findings indicate that while foundation models hold promise, pre-training remains vital for handling linguistic variations across languages for SER.**

***Index Terms*—cross-language, speech emotion recognition, foundation models, transformers, multilingual data, and self-supervised learning.**

I. Introduction

Speech emotion recognition (SER) is a technique for understanding human communication, both interpersonal and between people and machines [1]. With the potential to be a technology in forthcoming artificial general intelligence, new applications of SER are emerging at a pace, ranging from healthcare to transportation, forensics to education, entertainment to social media. SER classifies emotional categories by analysing audio signals such as pitch, intensity, and spectrograms [2]. However, dealing with cross-lingual inputs makes the job difficult since slight cultural and linguistic variances cast suspicion on the performance of the SER systems [3]. Though emotional expressions are universal across languages, the peculiarities of each language present an impediment to emotional interpretation. Therefore, advancing the capabilities of SER necessitates an understanding of the interactions between language, culture, and affective expression [4]. Machine learning (ML)-enabled SER systems have outperformed legacy emotion recognition systems and are now gaining traction in industry and academia [5]. As SER systems are now capable of solving the riddle of understanding and modelling human emotion with the aid of various context variables such as gender, age, dialect, and culture, it is now imperative to incorporate foundation model-like abilities into SER systems [1], [2], [6]. It would allow SER systems to understand cross-lingual emotion and act as a bridge towards the revolution in human-machine interaction (HCI) by enabling effective service delivery in a wider range of real-world applications.

The performance of ML-based SER in cross-language scenarios faces limitations due to several factors. A key factor is the ***language and cultural barrier*** - while human emotions are universal, their expression differs based on one's language, speech patterns and culture. As a result, the same emotion manifests with diverse cues and syntax across languages, posing a challenge [7], [8]. Another complication arises from ***cultural-specific linguistic nuances*** affecting emotional cues. Datasets used to train ML-based SER capture the nuances of a single language/culture. While performance may be adequate in that language, understanding other speech patterns faces issues [9]. Additionally, ***overlapping phonemes*** can cause misclassification exacerbated by phonetic relatedness between languages. SER requires extensive ***annotated data*** to learn emotion patterns, but datasets are limited for many languages. Without sufficient data, classification accuracy is restricted [10]. Current SER is also influenced by a single language. Addressing ***biases*** through multi-language training with fairness could help. However, variability in emotion expression and data scarcity present hurdles [11].

Recently, large foundation models pre-trained on massive corpora using self-supervised techniques have achieved results across natural language understanding tasks by learning robust cross-lingual representations [12], [13]. Their ability to understand relationships between languages without direct translation opens up possibilities for more universally applicable multilingual models [14]. Their scale of training allows them to discover deeper mappings between languages than was previously possible. Where traditional models focus narrowly on attributes of a single language, foundation models can learn shared semantic structures underlying emotional expressions across diverse cultures and tongues [9]. This expanded scope of understanding makes them suited to bridge gaps faced by traditional cross-language SER systems. By leveraging relationships between languages at both semantic and phonetic levels, foundation models hold the promise of speech emotion classifiers that generalise more effectively across linguistic barriers [15]. Foundation models enabling cheaper annotation may help generate annotated datasets to tackle the lack of data and aid effective cross-language SER design [16].

In this paper, we evaluate multiple multilingual foundation models - Wav2Vec2 [17], XLSR [18], Whisper [19], and MMS [20] - on benchmark cross-language SER datasets. Specifically, we analyse their ability to bridge the language and cultural gap when classifying emotions from speech data in multiple source and target languages. To the best of our knowledge, this represents the first study that assesses the capabilities of these foundation models for advancing the field of cross-language SER. The results from our experiments aim to provide insights into whether and how these models are able to learn cross-lingual speech representations that generalise better for emotion classification across languages compared to traditional approaches.

The ***major contributions of this work*** include:

- We experimentally investigated the possibility of using foundation models (Wav2Vec2 [17], XLSR [18], Whisper [19], and MMS [20]) to bridge the gap in cross-language SER.
- We used four different language corpora for speech emotion detection and evaluated the performance of pre-trained foundation models. Here we also note that the proposed method is scalable to many languages.
- We also provided the generalisation and robustness performance of cross-language SER under noisy data conditions. We also evaluated the few-shots adaptation performance of the cross-language SER. We further reported the performance of cross-language SER with the language information available at the pre-training of the foundation models.

The next section covers the related work II. Section III describes the models and datasets utilised in the research. Section IV focuses on the experiments and their results. Section V closes the paper and provides a way forward.

II. Related Work

This section provides a concise review of related work, including cross-lingual emotion recognition, multimodal emotion recognition, transformer-based emotion recognition approaches, and foundation models for SER.

A. Cross-Language SER

Cross-language emotion recognition aims to identify emotions in speech data across different languages and domains [3], [21], [22]. A key challenge is the limited availability of labelled data for low-resource languages like Urdu [3], Persian [22], or Marathi [21]. Additionally, using emotion recognition models trained on a single language or corpus limits generalizability due to domain mismatch between datasets [23]. Prior work has proposed various techniques to address these issues. Feature selection methods identify relevant features to represent emotions across modalities [24], [25]. Domain adaptation reduces distribution discrepancies between source and target domains via adversarial learning [26]). Data augmentation expands training data through transformations like speech synthesis and pitch shifting [27]–[29]. Multimodal fusion combines speech and text using attention, tensors, or graph networks [30]–[33]. Evaluation on benchmark datasets demonstrates the effectiveness of these approaches [3], [21]–[24], [26], [27], [34]–[36]. This prior work lays the groundwork for developing more generalised cross-language emotion recognition.

B. Transformers in Emotion Recognition

Transformers have advantages over RNNs and CNNs for emotion recognition tasks due to their ability to model long-range dependencies and perform parallel computation [2], [37]). They can effectively harness semantic and acoustic information from speech data to capture interactions between modalities such as audio and text [38]. Several studies have applied transformers for emotion recognition. Chen et al. [38] developed a key-sparse transformer for multimodal emotion classification focusing on emotion-related information. Wagner et al. [9] analysed the impact of model size and pre-training data on transformer performance, finding larger models pre-trained on more diverse data improved emotion prediction. Zenkov et al. [39] integrated CNNs with a transformer encoder to classify emotions from the RAVDESS dataset. Li et al. [40] proposed a multi-head self-attention-based transformer achieving results on IEMOCAP [34], MSP-Podcast [41], and MOSI [36] datasets. Triantafyllopoulos et al. [42] demonstrated transformers are sensitive to sentiment and negation through probing emotion recognition models.

C. Pre-trained models for SER

Multimodal emotion recognition involves identifying human affective states from multiple sources of information, including audio, text, visual, etc. This approach has demonstrated superior accuracy compared to single-modality methods [43]. Nevertheless, there are several challenges associated with multimodal emotion recognition, such as feature extraction

difficulties [44], feature alignment complexities [45], fusion techniques [46], dealing with missing or noisy data [44]. Therefore, more advanced methods are needed to effectively exploit information from multimodal data and provide a richer understanding of human emotions.

Several methods have been proposed for multimodal emotion recognition to improve performance using pre-trained models for feature extraction. Makiuchi et al. [47] proposed a cross-representation speech model combining self-supervised features from audio and text features extracted with Transformer models, achieving state-of-the-art results on IEMOCAP using score fusion. Tang et al. [48] propose a feature fusion method for facial expression and speech using attention mechanisms, showing improved accuracy on RAVDESS. Yoon et al. [49] proposed a deep dual recurrent encoder model using text and audio simultaneously to better understand speech data, outperforming previous methods on IEMOCAP emotion classification. Few recent works have proposed novel fusion techniques using hybrid transformer models [38], [43]. The hybrid transformer models combine transformer architectures, such as encoder-decoder or encoder-only, for better multimodal performance [38]. For example, Chen et al. [38] propose a key-sparse attention model fusing data efficiently using an encoder-decoder transformer. Wagner et al. [43] proposed a progressive fusion model using an encoder-only transformer to fuse data through refined iterations, preserving modality information while enhancing cross-modality interactions.

D. Foundation Models for SER

Foundation models have shown potential for speech emotion recognition by learning representations from large unlabeled speech datasets. For example, von Neumann et al. [50] pretrained a foundation model called SpeechBERT on 13,000 hours of unlabeled speech data from podcasts, achieving strong zero-shot transfer capabilities for SER tasks. Bender et al. [51] analyzed foundation models trained on speech to understand what linguistic patterns they learn and how robust their representations are. In our previous work [15], recent advances in audio foundation models were covered. By leveraging large amounts of audio data, these models have demonstrated abilities in various audio tasks including automatic speech recognition (ASR), text-to-speech, and music generation. Notably, foundation models like Wav2Vec2 [17], XLSR [52], Whisper [19], MMS [20], SeamlessM4T [12] have started showing capabilities as universal translators for multiple speech tasks across up to 100 languages without task-specific systems. Latif et al. [15] also presented an analysis of state-of-the-art methodologies regarding foundation large audio models, their performance benchmarks, and their applicability to real-world scenarios. Current limitations are also highlighted and insights are provided into potential future research directions for large audio models with the intent to spark discussion and foster innovation in next-generation audio processing systems. Careful analysis of biases is also needed as foundation models are deployed for real-world SER systems [11]. While audio foundation models show abilities, this research area remains in the early development stages. Further exploration and advancements are needed to fully realise the capabilities of these large language models for audio and speech-related applications, including potential pathways such as improving SER systems.

III. Model Architectures and Datasets

In this section, we describe the foundation models and datasets used for fine-tuning. We have selected various foundation models for comparison to gauge their performance against the baseline CNN-LSTM model. Below, we provide an overview of these models and datasets.

A. Baseline Model

Our baseline model incorporates a convolutional encoder structure coupled with a Bidirectional LSTM (BLSTM) for classification tasks. The convolutional layers within the encoder are designed to capture high-level emotional features. In line with past research [53]–[55], we use a larger kernel size for the initial convolutional layer and subsequently a smaller kernel for the subsequent layers. The encoder's features are then passed to the BLSTM layer, housing 128 LSTM units, to capture emotional contexts. These outputs from BLSTM are fed into a dense layer consisting of 128 units, generating discriminative features for the subsequent softmax layer. Overall, the model is trained using the cross-entropy loss for categorical SER.

B. Pretrained foundation models

We employ a simple head architecture and build it on top of established foundational models. Among the chosen foundation models for pre-training, we opt for esteemed models such as Wav2Vec2 [17], XLSR [52], Whisper [19], and MMS [20]. These models gain recognition for their training on vast, multilingual datasets. A comprehensive overview of these models, focusing on their scale, the datasets they train on, and the range of languages encompassed in their training data, is provided in Table I. We employ multilingual foundational models and fine-tune them for cross-language SER. By doing so, we contrast their capabilities against a conventional CNN-LSTM baseline, aiming to discern the effectiveness of these models in bridging the gap in cross-language emotion detection in speech.

Wav2Vec2 [17], a self-supervised model that learns by masking speech input in the latent space and tackling a contrastive task based on quantized latent representations. It was pre-trained using the Librispeech (LS-960) dataset, which lacks transcriptions and consists of 960 hours of audio. Additionally, they incorporated speech data from LibriVox (LV-60k). Notably, on the clear 100-hour segment of the Librispeech dataset, Wav2Vec2 outperformed previous benchmarks by only using 1% of the typically required labelled data.

XLSR, as introduced by Conneau et al. [52], stands as a pivotal model in the domain of cross-lingual speech representation learning. The foundation of XLSR is its pretraining on raw speech waveforms from a diverse array of languages. This approach is an extension of Wav2Vec2 but with a specific

TABLE I: Details on pre-trained foundation models, dataset, and a number of languages.

Model	Alias	Dataset	Hours	Languages
Wav2Vec2-base	Wav2Vec2	LibriSpeech	960	English
XLSR	XLSR	Common Voice BABEL Multilingual LibriSpeech (MLS)	50k+	50+
Massively Multilingual Speech	MMS	MMS-lab(44.7K hours) MMS-unlab(7.7K) MMS-lab-U(55K hours)	107k	1000+
Whisper	Whisper	Multitask training data (680k hours)	680k	96+

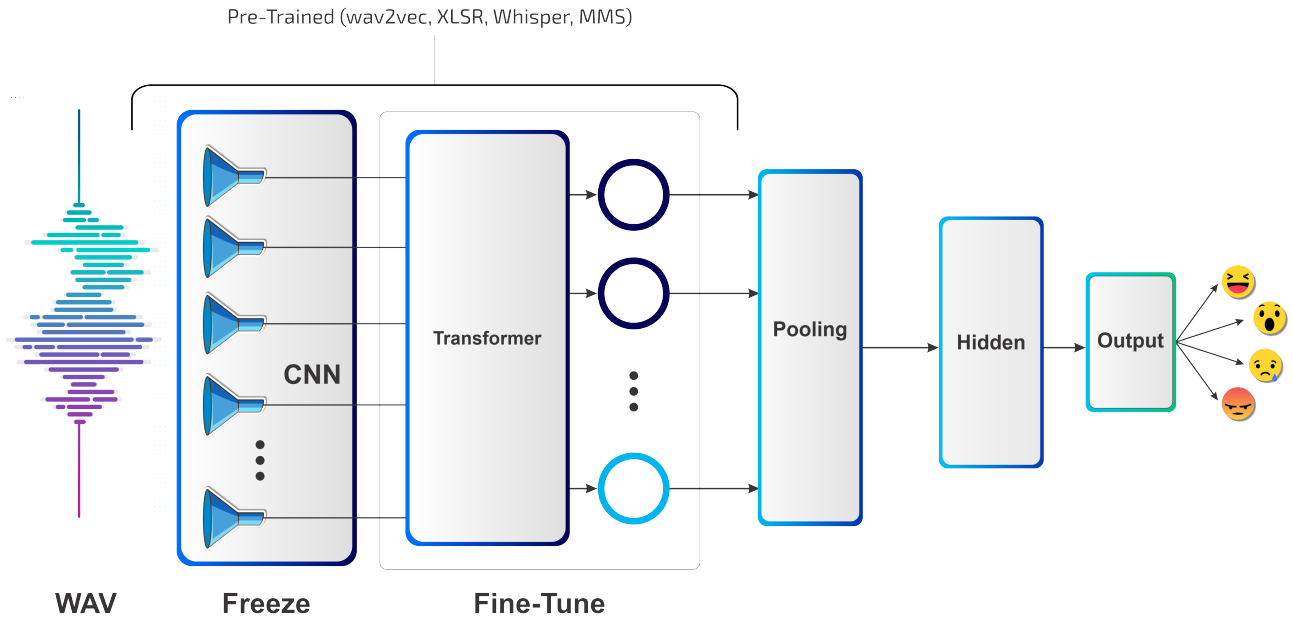


Fig. 1: Architecture build on top of W2V2/XLSR/Whiper/MMS

focus on cross-lingual settings. The pretraining phase involves solving a contrastive task that matches masked feature encoder outputs. The datasets that provided the bedrock for this expansive pretraining are Common Voice, BABEL, and Multilingual LibriSpeech (MLS). This comprehensive pretraining strategy not only boosts the model's ability to recognize and understand different languages but also sets the stage for effective fine-tuning. When subsequently tuned for specific tasks, XLSR demonstrates the ability to rival models that are individually optimized for each language.

Whisper [19] is trained through weakly supervised learning objectives. These objectives include tasks like Voice Activity Detection (VAD), language detection, and Automatic Speech Recognition (ASR), among others. The innovative facet of Whisper lies in its training methodology. By employing a colossal supervised dataset that spans over 680,000 hours of labelled audio data, it pushes the boundaries of weakly supervised speech recognition. Furthermore, the model underscores the potency of zero-shot transfer as a mechanism to significantly bolster the robustness of speech detection systems.

Massively Multilingual Speech (MMS) was introduced by Pratap et al. [20]. This initiative aimed to significantly expand the range of supported languages in speech technology by a notable 10-40x, depending on specific speech tasks. Central to their approach was the effective use of self-supervised learning. They curated a labelled dataset, MMS-lab, encompassing speech audio from 1,107 languages, totalling 44.7K hours. In parallel, they assembled an unlabelled dataset, MMS-unlab, with audio recordings without associated text from 3,809 languages, amounting to 7.7K hours. Additionally, an unlabelled variant of MMS-lab, designed for pre-training and language identification, named MMS-lab-U, spanned 1,362 languages and contributed 55K hours. With these resources, they developed a speech system capable of supporting a language count ranging from 1,100 to a vast 4,000.

For fine-tuning these models, we follow [9], [56] and the parameter-efficient fine-tuning (PEFT) technique. Regardless of the fine-tuning method employed, we also make sure of the consistency in our downstream cross-language SER architecture. To encapsulate, our efforts spanned from maintaining the original state of foundation models, adapter tuning, and modifying the embedding prompt, to using Lowrank approximation (LoRa) [57]. We implement average pooling on the

TABLE II: Results UAR (%) of cross-language experiments evaluating the performance of various pre-trained foundation models.

Source	Target	Models performance UAR (%)				
		CNN-LSTM	Wav2Vec2	XLSR	Whisper	MMS
IEMOCAP	EMODB	32.02	34.28	35.31	36.53	37.81
	EMOVO	56.58	60.89	62.20	63.22	63.81
	URDU	46.68	49.12	52.25	52.17	53.03
EMODB	IEMPCAP	40.10	43.30	44.15	44.03	45.85
	EMOVO	43.23	46.22	48.15	48.10	48.22
	URDU	54.15	56.26	59.01	59.52	59.13
URDU	IEMPCAP	38.92	40.26	42.54	43.89	44.08
	EMODB	51.82	53.12	55.22	56.85	56.51
	EMOVO	45.28	48.13	49.82	50.02	50.25
EMOVO	IEMPCAP	48.81	50.82	51.27	52.32	51.53
	EMODB	53.21	54.32	56.60	57.20	57.52
	URDU	56.52	57.21	60.81	60.21	61.10

hidden states from the final transformer layer, followed by processing through a hidden and then an output layer. For the downstream task fine-tuning, the ADAM optimiser is used alongside the cross-entropy loss, a commonly utilised loss function for classification. Our chosen learning rate is set at $1e-4$. The training lasts for 5 epochs with a batch size of 16, and we retain the model checkpoint that showcases the best results on the development set.

C. Datasets

To broaden the scope of our findings, we chose publicly accessible datasets representing four different languages. These corpora were selected due to their availability and to incorporate linguistic variety into our evaluations. An overview of each data collection is provided below.

1) IEMOCAP (English)

The IEMOCAP corpus, cited in [34], is a widely-used public collection of multimodal emotional data in English. Different annotators labelled the utterances in categorical and dimensional labelling schemes. Based on previous research [58], [59], we focused on four emotions: angry, sad, happy, and neutral. These emotions represented 5531 samples in the IEMOCAP dataset.

2) EMODB (German)

The EMODB corpus [60] is a well-known German emotional speech dataset. It features ten professional speakers conveying seven varied emotions through ten German sentences. In our research, we utilise 420 utterances: 127 angry, 143 sad, 79 neutral, and 71 happy expressions. This selection supported our detailed evaluation of cross-language emotion recognition.

3) EMOVO (Italian)

The EMOVO corpus [61], is an Italian emotional speech dataset. It includes 14 sentences, each delivered by six actors—three males and three females—expressing seven distinct emotions: anger, disgust, fear, joy, sadness, surprise, and neutral. In our research, we focused on 336 utterances that fit into four emotions: angry, happy, neutral, and sad, with each emotion having 84 utterances. This dataset is used in conducting a thorough cross-language SER evaluation.

4) URDU (Urdu)

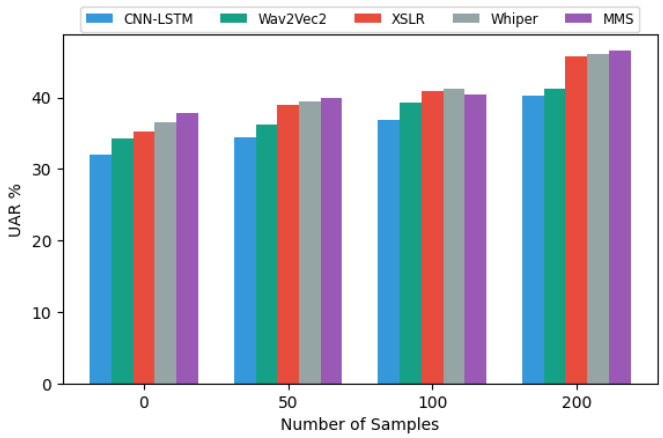
The URDU dataset [3], is an emotional speech collection in the Urdu language. It encompasses a total of 400 utterances, each reflecting one of the four fundamental emotions: angry, happy, neutral, and sad. This dataset features recordings from 38 distinct speakers, with 27 males and 11 females, all of whom were sourced from various Urdu talk shows available on YouTube. For our study, we've incorporated all 400 utterances, ensuring an equal representation of each emotion with 100 utterances each.

IV. Experiments and Results

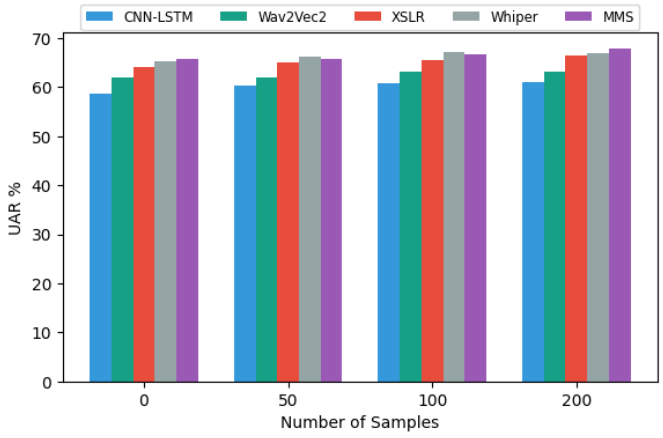
In this section, we evaluate and display the outcomes using various models. We employ these models to categorize emotions and gauge their effectiveness using the unweighted average recall (UAR). UAR is a popular metric in emotion recognition as it provides a balanced score, especially when the data for certain emotions might be imbalanced compared to others. We conducted each experiment five times and presented the average UAR for all results. Throughout our tests, we adhere to a speaker-independent SER approach.

A. Benchmarking Results

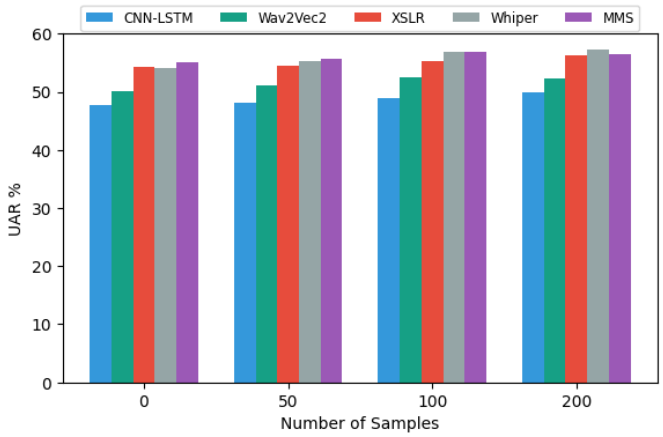
In this study, we conduct cross-language SER, training our model on source data and then evaluating its performance on unseen target data in a different language. We utilise datasets from four different languages including English, German, Italian, and Urdu. Our focus narrows to four primary emotional states: happy, sad, neutral, and angry. For experiments, we set out to see how these multilingual foundation models stack up against models like Wav2Vec2, which is solely pre-trained on English data, as well as the baseline CNN-LSTM model. Starting our experiment, we fine-tuned models using the IEMOCAP dataset and evaluated them on other language datasets. Those models that have been pre-trained with considerable data volume stand out, surpassing conventional architectures baseline CNN-LSTM. To make our observations generalisable, we perform multiple evaluation strategies across the selected four datasets and the results are presented in Table II. Results show the dominant position of foundation models over the conventional CNN-LSTM methods in the field of



(a) IEMOCAP to EMODB

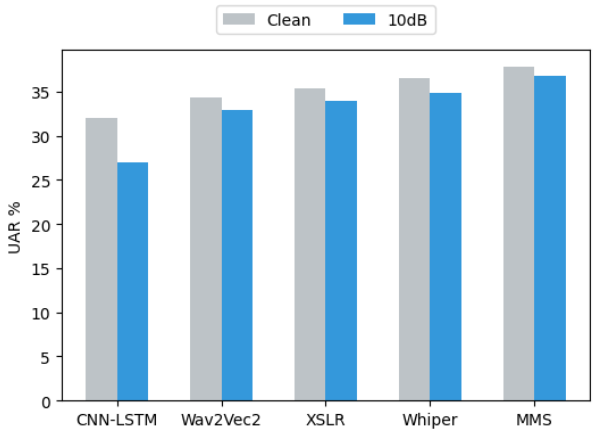


(b) IEMOCAP to EMOVO

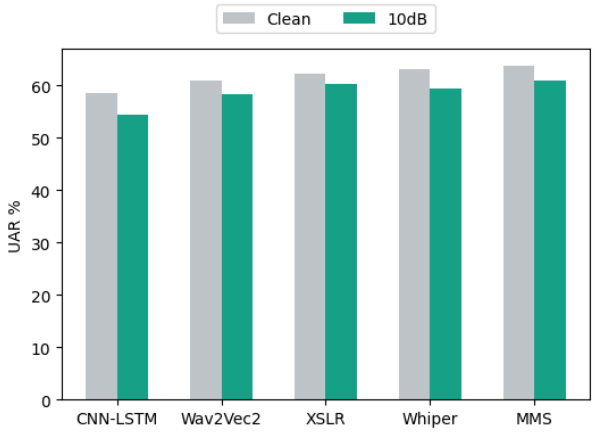


(c) IEMOCAP to URDU

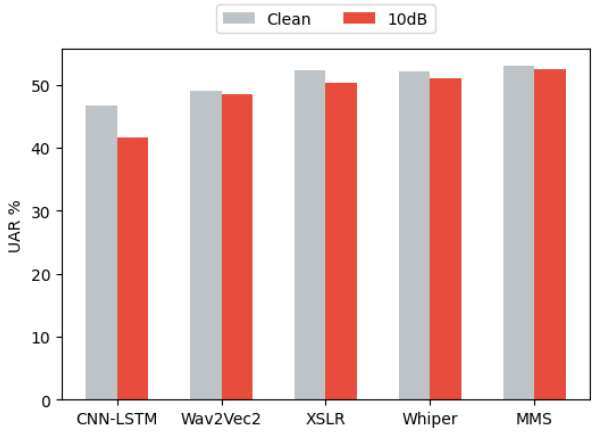
Fig. 2: Cross-language SER performance comparison of CNN-LSTM and various pre-trained foundation models across three datasets (EMODB, EMOVO, URDU) for varying sample sizes, as measured by UAR(%). The models are fine-tuned on IEMOCAP and evaluated on the target datasets.



(a) IEMOCAP to EMODB



(b) IEMOCAP to EMOVO



(c) IEMOCAP to URDU

Fig. 3: Cross-language SER performance (UAR%) comparison of CNN-LSTM and pre-trained foundation models on clean speech vs noisy data (10dB SNR) from the target datasets. The models are fine-tuned on IEMOCAP and evaluated on clean and noisy versions of EMODB, EMOVO and URDU datasets.

cross-language SER. This shows that the models with a diverse linguistic background tend to perform better in cross-language SER tasks compared to ones like Wav2Vec2 which is pre-trained on English data. Our findings in cross-language SER

underscore the advantages of foundation models pre-trained on rich and varied linguistic datasets. Such extensive pre-training evidently positions these models favourably for cross-language SER tasks, as illustrated in Table II.

B. Few-shots Adaptation

In this experiment, we delve into the impact of few-shot adaptation for cross-language settings. Essentially, we combine a subset of samples from a target language with our main training data to observe the outcomes. The IEMOCAP dataset serves as our primary training resource in this experiment, and we use other datasets as test data. We fine-tuned foundation models and the results of our findings are illustrated in Figure 2. To cover a wide spectrum, we alter the sample size, beginning with zero and ramping up to 200. We initiate our tests with 50 samples for each individual emotion. From there, we methodically increase the sample count, examining results at 100, 150, and 200 sample thresholds. As depicted in Figure 2, a clear pattern emerges: employing few-shot adaptation while fine-tuning the pre-trained foundation model notably boosts cross-language SER, surpassing the conventional models like the CNN-LSTM model. Importantly, this experiment also highlights the elevated efficiency of foundation models pre-trained on datasets comprising multiple language data, such as XLSR, Whiper, and MMS. These models outperform Wav2Vec2, which is only trained on English language data. This result highlights how using varied language data for training can make a difference, especially in recognising emotions across languages. Incorporating a few samples from the target data into the training set can notably boost performance, paving the way for real-world applications of SER.

C. Robustness of Pre-trained Models

In this experiment, we delve into the robustness of various architectures, especially contrasting traditional models like CNN-LSTM with transformer-based models pre-trained on vast and diverse datasets. The test conditions involve the introduction of ambient noise—specifically kitchen, park, station, traffic, and cafeteria sounds from the DEMAND dataset [62]. This noise is interspersed randomly within the test dataset. Performance assessments are then made on the data with a signal-to-noise (SNR) ratio of 10 dB, against clean speech, and the findings are captured in Figure 3. Several observations emerge from this analysis. Pre-trained foundation model, given their extensive training on a large corpus of data, seemingly display an innate ability to handle noisy disruptions better. It is plausible that their expansive training data encompassed various noisy environments, furnishing them with the capability to better adapt to, and process, distorted auditory signals. Their SER performance, in the context of noise tolerance, distinctively eclipses that of the conventional CNN-LSTM model.

Furthermore, it becomes evident that sheer volume and diversity in training data play pivotal roles in noise resilience. Models like XLSR, Whiper, and MMS, pre-trained on substantial multilingual datasets, illustrate superior performance metrics compared to the Wav2Vec2 base. This differential is not just attributable to the advanced transformer architecture but also the breadth of their training data. Specifically, the Wav2Vec2 base model, constrained by its training solely on English data, struggles to match the versatility and adaptability of its more extensively trained counterparts. This reaffirms the notion that diversity in training—both in terms of language and acoustic conditions—equips models with a more holistic, noise-resistant capability.

V. Conclusions and Outlook

In this paper, we evaluated the performance of different foundation models for cross-language speech emotion recognition. Based on our experiments and analysis, we conclude the following:

- Foundation models like XLSR, Whisper, and MMS significantly outperform traditional CNN-LSTM approaches for cross-language SER, achieving higher UAR scores across different language pairs. This establishes the superiority of foundation models in handling cross-lingual learning for emotion recognition.
- As found previously [9], Wav2Vec2, when initialised randomly, showed performance comparable to CNN-LSTM. However, models like XLSR, Whisper, and MMS, which are pre-trained on massive multi-language datasets, demonstrate improved performance in cross-language SER compared to Wav2Vec2 trained on single-language data. The distinct advantage underscores the significance of diverse pre-training datasets in elevating the capabilities of speech models for fine-tuning tasks.
- Adapting the foundation models with a few target language samples resulted in substantial gains over the baseline, demonstrating their ability to effectively leverage limited target data.
- The foundation models also exhibited better robustness over CNN-LSTM when evaluated on noisy target data, maintaining higher UAR scores.

In conclusion, while foundation models hold promise for cross-language tasks, self-supervised pre-training currently plays a vital complementary role in equipping them with the necessary skills for handling linguistic and cultural variations across languages. Further research can explore inductive biases that facilitate improved cross-lingual transfer ability of foundation models.

References

[1] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[2] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, "Survey of deep representation learning for speech emotion recognition," *IEEE Transactions on Affective Computing*, 2021.

[3] S. Latif, A. Qayyum, M. Usman, and 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.

[4] S. Latif, "Deep representation learning for speech emotion recognition," Ph.D. dissertation, University of Southern Queensland, 2022.

[5] M. J. Al-Dujaili and A. Ebrahimi-Moghadam, "Speech emotion recognition: a comprehensive survey," *Wireless Personal Communications*, vol. 129, no. 4, pp. 2525–2561, 2023.

[6] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019.

[7] S. Sitaram, K. R. Chandu, S. K. Rallabandi, and A. W. Black, "A survey of code-switched speech and language processing," *arXiv preprint arXiv:1904.00784*, 2019.

[8] G. Costantini, E. Parada-Cabaleiro, D. Casali, and V. Cesarini, "The emotion probe: On the universality of cross-linguistic and cross-gender speech emotion recognition via machine learning," *Sensors*, vol. 22, no. 7, p. 2461, 2022.

[9] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[10] J. de Lope and M. Graña, "An ongoing review of speech emotion recognition," *Neurocomputing*, vol. 528, pp. 1–11, 2023.

[11] S. Latif, H. S. Ali, M. Usama, R. Rana, B. Schuller, and J. Qadir, "Ai-based emotion recognition: Promise, peril, and prescriptions for prosocial path," *arXiv preprint arXiv:2211.07290*, 2022.

[12] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman *et al.*, "Seamlessm4t-massively multilingual & multimodal machine translation," *arXiv preprint arXiv:2308.11596*, 2023.

[13] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[14] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He *et al.*, "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt," *arXiv preprint arXiv:2302.09419*, 2023.

[15] S. Latif, M. Shoukat, F. Shamshad, M. Usama, H. Cuayáhuitl, and B. W. Schuller, "Sparks of large audio models: A survey and outlook," *arXiv preprint arXiv:2308.12792*, 2023.

[16] S. Latif, M. Usama, M. I. Malik, and B. W. Schuller, "Can large language models aid in annotating speech emotional data? uncovering new frontiers," *arXiv preprint arXiv:2307.06090*, 2023.

[17] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[18] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[20] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi *et al.*, "Scaling speech technology to 1,000+ languages," *arXiv preprint arXiv:2305.13516*, 2023.

[21] P. Lahoti, N. Mittal, and G. Singh, "A survey on nlp resources, tools, and techniques for marathi language processing," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 2, pp. 1–34, 2022.

[22] S. Deng, N. Zhang, Z. Sun, J. Chen, and 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*, vol. 34, no. 10, 2020, pp. 13 773–13 774.

[23] P. Godard, G. Adda, M. Adda-Decker, J. Benjumea, L. Besacier, J. Cooper-Leavitt, G.-N. Kouarata, L. Lamel, H. Maynard, M. Müller *et al.*, "A very low resource language speech corpus for computational language documentation experiments," *arXiv preprint arXiv:1710.03501*, 2017.

[24] T. Özseven, "A novel feature selection method for speech emotion recognition," *Applied Acoustics*, vol. 146, pp. 320–326, 2019.

[25] B. T. Atmaja, A. Sasou, and M. Akagi, "Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion," *Speech Communication*, 2022.

[26] S. Latif, J. Qadir, and M. Bilal, "Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 732–737.

[27] S. Kshirsagar and T. H. Falk, "Cross-language speech emotion recognition using bag-of-word representations, domain adaptation, and data augmentation," *Sensors*, vol. 22, no. 17, p. 6445, 2022.

[28] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Deep architecture enhancing robustness to noise, adversarial attacks, and cross-corpus setting for speech emotion recognition," *Proc. Interspeech 2020*, pp. 2327–2331, 2020.

[29] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, and S. Narayanan, "Data augmentation using gans for speech emotion recognition." in *Interspeech*, 2019, pp. 171–175.

[30] J. Liu, W. Zheng, Y. Zong, C. Lu, and C. Tang, "Cross-corpus speech emotion recognition based on deep domain-adaptive convolutional neural network," *IEICE TRANSACTIONS on Information and Systems*, vol. 103, no. 2, pp. 459–463, 2020.

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

[32] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI open*, vol. 1, pp. 57–81, 2020.

[33] G. Shen, R. Lai, R. Chen, Y. Zhang, K. Zhang, Q. Han, and H. Song, "Wise: Word-level interaction-based multimodal fusion for speech emotion recognition." in *Interspeech*, 2020, pp. 369–373.

[34] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[35] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[36] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.

[37] S. Latif, A. Zaidi, H. Cuayahuitl, F. Shamshad, M. Shoukat, and J. Qadir, "Transformers in speech processing: A survey," *arXiv preprint arXiv:2303.11607*, 2023.

[38] W. Chen, X. Xing, X. Xu, J. Yang, and J. Pang, "Key-sparse transformer for multimodal speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6897–6901.

[39] I. Zenkov, "Transformer-cnn emotion recognition," https://github.com/IliaZenkov/transformer-cnn-emotion-recognition, 2021.

[40] Y. Li, Z. Li, Z. Zhang, X. Li, and J. Li, "Speech emotion recognition transformer: A novel end-to-end model for ser," *Neurocomputing*, vol. 454, pp. 1–10, 2021.

[41] J. Park and C. Busso, "Msp-podcast: A large-scale dataset of natural and emotionally evocative speech," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 112–119.

[42] A. Triantafyllopoulos, J. Wagner, H. Wierstorf, M. Schmitt, U. Reichel, F. Eyben, F. Burkhardt, and B. W. Schuller, "Probing speech emotion recognition transformers for linguistic knowledge," in *Proc. Interspeech 2022*, 2022, pp. 146–150.

[43] Q. Wang, M. Wang, Y. Yang, and X. Zhang, "Multi-modal emotion recognition using eeg and speech signals," *Computers in Biology and Medicine*, vol. 149, p. 105907, 2022.

[44] D. Liu, Z. Wang, L. Wang, and L. Chen, "Multi-modal fusion emotion recognition method of speech expression based on deep learning," *Frontiers in Neurorobotics*, vol. 15, p. 697634, 2021.

[45] Y. Liu, H. Sun, W. Guan, Y. Xia, and Z. Zhao, "Multi-modal speech emotion recognition using self-attention mechanism and multi-scale fusion framework," *Speech Communication*, vol. 139, pp. 1–9, 2022.

[46] C.-P. Ho, C.-C. Yang, S. Kim, and Y.-N. Lee, "Multi-head attention fusion networks for multi-modal speech emotion recognition," *Computer Speech & Language*, vol. 65, p. 101122, 2020.

[47] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech emotion recognition using self-supervised features," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6922–6926.

[48] W. Tang, F. He, Y. Liu, and Y. Duan, "Matr: Multimodal medical image fusion via multiscale adaptive transformer," *IEEE Transactions on Image Processing*, vol. 31, pp. 5134–5149, 2022.
[49] D. Yoon, S. Lee, and H. Lee, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 112–118.
[50] Y.-S. Chuang, C.-L. Liu, H.-Y. Lee, and L.-s. Lee, "Speechbert: An audio-and-text jointly learned language model for end-to-end spoken question answering," *arXiv preprint arXiv:1910.11559*, 2019.
[51] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 610–623.
[52] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. Interspeech 2021*, 2021, pp. 2426–2430.
[53] D. Dai, Z. Wu, R. Li, X. Wu, J. Jia, and H. Meng, "Learning discriminative features from spectrograms using center loss for speech emotion recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7405–7409.
[54] J. Gideon, M. G. McInnis, and E. M. Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (addog)," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 1055–1068, 2019.
[55] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Multitask learning from augmented auxiliary data for improving speech emotion recognition," *IEEE Transactions on Affective Computing*, 2022.
[56] T. Feng and S. Narayanan, "Peft-ser: On the use of parameter efficient transfer learning approaches for speech emotion recognition using pre-trained speech models," *arXiv preprint arXiv:2306.05350*, 2023.
[57] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.
[58] S. Latif, M. Asim, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Augmenting generative adversarial networks for speech emotion recognition," *arXiv preprint arXiv:2005.08447*, 2020.
[59] P. Kumar, V. Kaushik, and B. Raman, "Towards the explainability of multimodal speech emotion recognition." in *Interspeech*, 2021, pp. 1748–1752.
[60] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.
[61] 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.
[62] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics ICA2013*, vol. 19, no. 1. Acoustical Society of America, 2013, p. 035081.