Interspeech 2024
1-5 September 2024, Kos, Greece

Exploring Multilingual Unseen Speaker Emotion Recognition: Leveraging Co-Attention Cues in Multitask Learning

Arnav Goel[1†#], *Medha Hira*[1†], *Anubha Gupta*[1,2]

[1]SBILab, Indraprastha Institute of Information Technology Delhi (IIIT-D), India
[2]MIRAE AI Systems Pvt. Ltd.

arnav21519,medha21265,anubha{@iiitd.ac.in}

Abstract

Advent of modern deep learning techniques has given rise to advancements in the field of Speech Emotion Recognition (SER). However, most systems prevalent in the field fail to generalize to speakers not seen during training. This study focuses on handling challenges of multilingual SER, specifically on unseen speakers. We introduce CAMuLeNet, a novel architecture leveraging co-attention based fusion and multitask learning to address this problem. Additionally, we benchmark pretrained encoders of Whisper, HuBERT, Wav2Vec2.0, and WavLM using 10-fold leave-speaker-out cross-validation on five existing multilingual benchmark datasets: IEMOCAP, RAVDESS, CREMA-D, EmoDB and CaFE and, release a novel dataset for SER on the Hindi language (BhavVani). CAMuLeNet shows an average improvement of approximately 8% over all benchmarks on unseen speakers determined by our cross-validation strategy.
Index Terms: Speaker emotion recognition, co-attention, multitask learning, new dataset

1. Introduction

In her seminal work on affective computing, Picard asserts that computers can achieve genuine intelligence and natural interactivity if we empower them with the ability to recognize and understand emotions [1]. Besides working with vision-based cues such as facial expressions and hand movements, humans excel in discerning emotions even when only auditory information is available [2]. This ability highlights the nuanced and adaptable nature of human emotional understanding, capable of discerning and processing emotions across a wide spectrum of speakers and voice types. Building on this premise, it is imperative to extend the capabilities of Speaker Emotion Recognition (SER) systems beyond speakers on which they have been trained.

Traditional SER systems rely on features including pitch, energy, MFCCs, and spectrograms for emotion recognition [3, 4]. With the emergence of deep learning methods, systems employing CNNs, Bi-Directional RNNs, and LSTMs are able to learn discernible features [5, 6, 7]. Transformer-based models, a more recent development, marked a significant advancement with the introduction of large pre-trained models (PTMs) trained under a self-supervised learning framework [8]. Recent advancements in weakly-supervised models, such as Whisper [9], which are trained on extensive corpora, have demonstrated superior performance on a diverse array of downstream tasks [10, 11, 12]. Attention mechanisms, including cross-attention [13], windowed-attention [14], and self-attention [15] along with multitask training have been explored to enhance the performance of SER [16, 17, 18]. Despite these advances, a critical challenge remains: the inability of modern SER systems to effectively adapt to unseen scenarios and speakers, resulting in performance that is inferior to human capabilities [19, 20].

This study contributes to the field by benchmarking various PTM embeddings in a transfer learning framework, specifically addressing unseen speaker recognition on five existing benchmark datasets and the 6th newly released dataset with this work. Although co-attention based fusion mechanisms have been used previously on the speaker emotion recognition downstream task for fusing features from multiple modalities [21] and multi-level acoustic information [22], their use on unseen speaker emotion recognition tasks along with multitask learning is yet to be thoroughly explored. This study addresses this gap by proposing an architecture that fuses features from the frequency domain and PTM embeddings.

Moreover, the variation in emotional expression across languages poses a distinctive challenge in multilingual SER, which is compounded by the scarcity of comprehensive datasets. To address this gap, we introduce a novel Hindi SER dataset, designed to enhance model training and benchmarking in Indian linguistic contexts. To the best of our knowledge, this is the *first open-source Hindi SER dataset*. Extending our efforts, we apply our methodology to French and German datasets, positioning our work as the first to benchmark Whisper's encoder in multilingual SER settings. The codes and dataset can be found on GitHub[1].

The key contributions of this work are three-fold:

1. We introduce BhavVani, the first-ever Hindi Speech Emotion Recognition dataset with over 9,000 utterances.
2. Our research uniquely benchmarks pre-trained model embeddings across six datasets in four languages (English, German, French, Hindi) for unseen speaker recognition, marking the first study of its kind to explore these embeddings, including Whisper, on this downstream task.
3. We introduce *CAMuLeNet*, a Co-Attention based Multitask Learning Network architecture that fuses frequency domain features with PTM features in a multitask framework of emotion and gender recognition, aiming to derive generalized representations for enhanced speaker emotion recognition.

2. Proposed Methodology

The proposed CAMuLeNet architecture described in Figure1, aims to fuse traditional frequency domain features of the spectrogram and MFCCs, with the features extracted from a pretrained Whisper encoder using the co-attention mechanism described next. We train this architecture through a multitask setup to improve performance on unseen speakers' emotion recognition.

[†]Equal contribution; [#]Corresponding Author

[1]https://github.com/arnav10goel/CAMuLeNet

10.21437/Interspeech.2024-1820

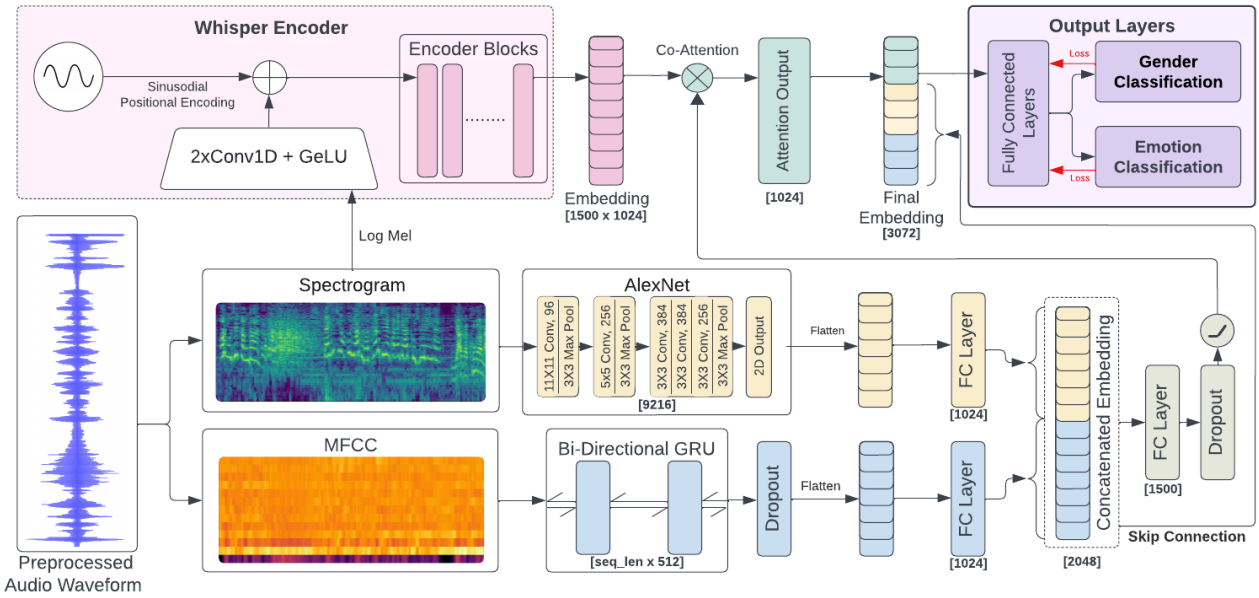


Figure 1: *CAMuLeNet: Co-Attention based Multitask Learning Network*

2.1. Extracting Features from the Frequency Domain

To capture frequency and pitch variations in audio clips for generalizing unseen speakers, we utilize frequency-domain features. The spectrogram and mel-frequency cepstrum coefficients (MFCCs) are represented as x_s and x_m, respectively, choosing 40 MFCCs. Both are used in their two-dimensional form, capturing time-frequency representation without average pooling. An audio clip x is preprocessed through padding and filtering for consistent sequence length, followed by feature extraction via an AlexNet encoder [23], treating the spectrogram as an image. The latent embeddings from AlexNet is a one-dimensional feature vector x'_s of size 4096. Concurrently, MFCCs undergo processing through a Bidirectional Gated Recurrent Unit (Bi-GRU) with two 256-sized hidden layers and a 0.2 dropout, producing a one-dimensional embedding x'_m sized (seq_len $\times$512).

2.2. Extracting Features through Transfer Learning

OpenAI's Whisper[2], a multilingual encoder-decoder model, exhibits state-of-the-art performance across various speech-to-text benchmarks. Trained on a vast and diverse audio dataset, it significantly improves speech recognition and translation tasks. We hypothesize that Whisper's encoder produces rich latent representations of audio samples. We pass the preprocessed audio clip x through the Whisper Encoder (which converts it into a mel spectrogram for processing), obtaining a 2D latent representation $x_w \in \mathbb{R}^{L\times W}$. We refrain from using pooling on this 2D representation to maintain the time-frequency information of these embeddings.

2.3. Co-Attention based Fusion

The three embeddings hold vital knowledge representation that we aim to fuse for improved performance. First, the derived embeddings from the spectrogram x'_s and MFCC features x'_m are passed through a fully connected (FC) layer with the same output dimension:

$$x_{m_{att}} = f_{att_m}(x'_m) \quad , \quad x_{s_{att}} = f_{att_s}(x'_s) \tag{1}$$

where $x_{s_{att}} \in \mathbb{R}^{1\times T}$, $x_{m_{att}} \in \mathbb{R}^{1\times T}$, and $T(= 1024)$ is a hyperparameter. The two transformed one-dimensional embeddings are concatenated to create a one-dimensional vector and passed through another FC layer:

$$x_{sm_{att}} = f_{att_{ms}}(x_{m_{att}} \oplus x_{s_{att}}) \tag{2}$$

[2]https://github.com/openai/whisper

where $x_{sm_{att}} \in \mathbb{R}^{1\times L}$. The activated concatenated features from (2) are sent through a FC layer activated by *ReLU* and sent through a *Layer Norm*. This output is multiplied with Whisper features x_w to generate attention-weighted Whisper features

$$x_{w_{att}} = f_{att_{ms}}(x_{sm_{att}} \otimes x_w), \tag{3}$$

where $x_{w_{att}} \in \mathbb{R}^{1\times W}$. Attention-weighted features are sent through a 3-layer network with 0.15 dropout, an FC layer with ReLU activation, and a terminal Layer Norm. These processed features, along with activated features from (2) sent via a skip-connection, are concatenated to form the network's final embedding for downstream classification. This co-attention mechanism attends frequency domain information with frame-level features from the Whisper encoder.

2.4. Multitask Learning

Utilizing multitask learning, our network concurrently trains on emotion and gender recognition, leveraging the significant influence of speaker gender on emotion. This approach aims to develop nuanced latent representations that captures intricate emotion-speaker correlations. We use categorical cross entropy loss $\mathcal{L}_{cce}$ for multiclass emotion recognition and binary cross entropy loss $\mathcal{L}_{bce}$ for gender recognition, reflecting the prevalent binary gender annotation. The combined multitask objective is defined as

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{cce} + \beta \cdot \mathcal{L}_{bce} + \gamma, \tag{4}$$

with α and β as tunable weights for each task and γ as a tunable parameter providing stability while training. Experimentally, setting α three times higher than β effectively balances training across tasks, attributed to the class count ratio and the relative simplicity of emotion recognition, thus imposing a higher penalty for better learning from the latter.

3. Datasets

3.1. Benchmark Datasets

This study utilizes five widely-used multilingual datasets (English, German, French) to assess pre-trained model embeddings and our methodology for unseen speaker emotion recognition. They are summarised next.

The CREMA-D dataset [24], featuring 7,442 clips from 91 voice actors (48 males, 43 females), offers recordings of 12 sentences expressed in 6 emotions in English. The IEMOCAP dataset [25] includes 10,039 clips from 10 actors (5 males, 5

females) in both scripted and spontaneous conversations, annotated with 4 emotion classes. The RAVDESS [26] dataset comprises 1,440 recordings from 24 actors (12 males, 12 females), each articulating two sentences across 8 emotional states. Each recording features one speaker expressing a single emotion, with equal distribution of male and female voices across the dataset. Among these, CREMA-D stands out with the largest variety of unique speakers from various ethnicity and is the largest corpus of speech samples among the existing datasets.

We utilize EmoDB [27] and CaFE [28] datasets for German and French languages, respectively. EmoDB includes 535 recordings from 10 actors (5 males, 5 females), featuring 10 sentences in *7 emotional states*. CaFE provides a collection of 936 clips from 12 actors (6 males, 6 females), each expressing 6 sentences across *7 emotions*.

3.2. Our Novel Hindi SER Dataset: *BhavVani*

In response to the under-representation of Indic languages in Speech Emotion Recognition (SER) and their absence in multilingual benchmarks such as SERAB [29], we release BhavVani, a novel dataset tailored for Hindi SER. This initiative aims to enrich the SER research landscape by incorporating the linguistic and cultural diversity of Indian languages, leveraging their morphological richness and unique emotional expressiveness.

3.2.1. Data Collection

The BhavVani dataset comprises approximately *13 hours of audio* across *8734* utterances, with an average clip length of 5.08 seconds. The audio clips are curated from the popular Indian sitcom "Sarabhai vs Sarabhai", sourced from prior work that used the show's text for sarcasm detection tasks [30]. To our knowledge, BhavVani is the *first open-source Hindi dataset tailored for speaker emotion recognition*. The dataset will be released to the community for further work. Each utterance is a single-speaker dialogue, annotated into one of seven categories: Neutral, Surprise, Enjoyment, Disgust and Anger, Fear, Sadness based off Ekman's seven basic human emotions [31].

3.2.2. Data Annotation and Validation

The BhavVani dataset was annotated by 18 native Hindi speakers, who received preliminary training to ensure annotation consistency and accuracy. Each annotator evaluated approximately 750 audio clips, identifying the speaker's gender, name, and perceived emotion while excluding clips with excessive background noise or multiple speakers. To uphold the integrity of the dataset, a rigorous validation procedure was followed involving three independent annotators reviewing each annotation. Annotations that were agreed by all three annotators, were marked as confirmed, while those with disagreements were re-evaluated, if necessary, and an alternative emotion was suggested. This meticulous annotation and validation method, evidenced by a *Fleiss Kappa score of 0.637*, highlights substantial annotator agreement, ensuring the reliability and quality of the dataset.

3.2.3. Dataset Statistics

The pie-charts in Figure 2 provide insights of the thoroughly annotated and curated dataset that had no missing labels. The dataset is balanced with male and female speakers. The gender has been annotated for all speakers, while the speaker name has been annotated primarily for the main characters. None of the other characters had an individual occurrence greater than 41, hence they were positioned under the umbrella of *'Others'*.

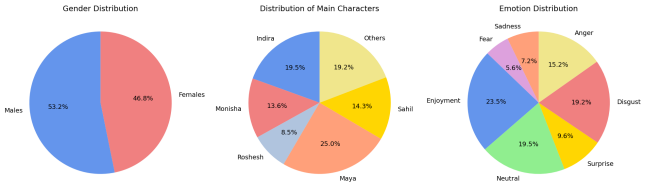


Figure 2: *BhavVani Dataset Statistics*

4. Experiments

4.1. Baseline Pre-Trained Models Employed

We employed encoders from pre-trained models (PTMs) such as Wav2Vec2.0 [32], WavLM [33], HuBERT [34], and Whisper [9] as baselines via transfer learning. Initially trained for ASR tasks, Wav2Vec2.0, WavLM, and HuBERT are transformer-based models with 95M parameters each trained through self-supervised learning on English datasets. We fine-tuned these models for multilingual SER using CTC loss. Whisper, distinguished as an encoder-decoder model, is trained with weak supervision on a diverse 680000-hour multilingual dataset, with our focus on its Base (74M) and Medium (769M) variants. We adapt Whisper models for SER by converting their 2D time-frequency outputs into 1D embeddings via average pool. Chosen for their excellence in speech processing, their potential with unseen speakers is yet to be fully explored.

4.2. Baseline Experiment Setup

In the baseline experiment, we provided 1D-embedding obtained in Section 4.1 as input to a CNN based feature extractor comprising of a 1D convolutional layer, batch normalization, ReLU activation, dropout (0.3), and max pool followed by flattening and two fully connected (FC) layers for classification. We utilized cross-entropy loss and Adam optimizer with a learning rate of 10^{-4} as higher rates led to model overshoot, rapid loss increase, and early underfitting within a few epochs. We trained the models on a NVIDIA A5000 GPU. Training was capped at 20 epochs with early stopping based on validation loss to prevent overfitting. For evaluating performance on unseen speakers, we followed 10-fold leave-speaker-out cross validation, wherein each dataset was segmented into 10 folds with each fold containing unique speakers. Thus, CREMA-D had around 9 new speakers at the time of validation that the model had not seen at the training time.

4.3. CAMuLeNet Training Setup

We extracted MFCC and spectrogram features from the pre-processed audio waveform. The spectrogram is derived by applying a Hamming window based short-time Fourier transform, with a window length of 40, a hop length of 10, setting the size of the FFT window to 800. The calculation of MFCCs involved generating 40 MFCC values with a hop length of 160, ensuring that these coefficients were compatible with the Hidden Markov Model Toolkit (HTK). The training of our model architecture was conducted using a NVIDIA A5000 GPU using batches of 64, leveraging the Adam optimizer with a learning rate of 5×10^{-5}. Dropout was maintained at 0.15 throughout the network. In the context of multitask learning, we determined experimentally that the model achieved optimal stability across different datasets when the weighting factors were set to $\alpha = 0.4$, $\beta = 0.1$ and $\gamma = 0.2$. These values, however, may require tuning to accommodate variations in dataset characteristics and training configurations. Remaining conditions are as described in Section-4.2.

Table 1: *Comparison of Baseline Methods and CAMuLeNet across various datasets*

PTM	CREMA-D		IEMOCAP		RAVDESS		EmoDB		CaFE		BhavVani	
	WA (↑)	WF1 (↑)	WA (↑)	WF1 (↑)	WA (↑)	WF1 (↑)	WA (↑)	WF1 (↑)	WA (↑)	WF1 (↑)	WA (↑)	WF1 (↑)
Wav2Vec2.0	0.451	0.409	0.427	0.408	0.396	0.377	0.469	0.455	0.417	0.372	0.251	0.317
HuBERT	0.529	0.499	0.461	0.458	0.587	0.546	0.486	0.479	0.424	0.397	0.253	0.332
WavLM	0.587	0.547	0.512	0.475	0.654	0.612	0.806	0.781	0.444	0.402	0.303	0.305
Whisper-Base	0.603	0.591	0.573	0.588	0.593	0.587	0.743	0.732	0.452	0.441	0.261	0.265
Whisper-Medium	**0.666**	**0.649**	**0.648**	**0.667**	**0.713**	**0.735**	**0.831**	**0.799**	**0.623**	**0.551**	**0.412**	**0.439**
Architecture	**Experiments on Co-Attention and Multitask Learning**											
CAMuLeNet (Ours)	**0.762**	**0.768**	**0.734**	**0.728**	**0.823**	**0.826**	**0.862**	**0.847**	**0.709**	**0.691**	**0.453**	**0.441**
	Ablation Study											
Ours (w/o Multitask)	0.719	0.721	0.703	0.697	0.782	0.784	0.853	0.844	0.672	0.683	0.431	0.429
Ours (w/o Co-Att & Multitask)	0.671	0.653	0.659	0.662	0.721	0.728	0.817	0.793	0.605	0.593	0.407	0.398

Note: WA stands for weighted accuracy and WF1 stands for weighted F1 score. These measures account for class imbalance.

5. Results and Discussion

This Section details our comparative analysis of baseline and proposed methods from 4.2, quantified by Weighted Accuracy (WA) and Weighted F1 score (WF1) across six datasets, presented in Table 1, complemented by an ablation study discussed later. We report mean metric values averaged over ten folds.

5.1. Analysis of Baseline Transfer Learning Results

As per Table-1, Whisper-Medium shows a huge performance jump over self-supervised based PTMs and its base architecture. The increase is noteworthy on the IEMOCAP (13% over WavLM), CaFE (18% over Whisper-Base) and BhavVani (14% over WavLM). The increase is higher on French language dataset due to Whisper being trained on a lot of audio chunks from the French language. Objective performance, however, remains sub-optimal on our Hindi SER dataset indicating the need for interventions to create more robust models and resources for Indic languages.

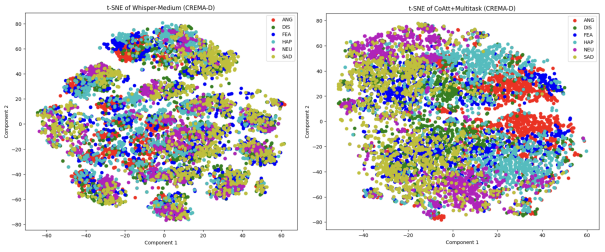
5.2. Analysis of Results on CAMuLeNet

The baseline analysis indicated Whisper-Medium as an optimal pre-trained embedding for our co-attention fusion method. As detailed in Section 4.3, we established a multi-task training framework and observed improved performance across all benchmarks. The introduction of the parameter γ into our loss function contributed to stabilized training, evidenced by a consistent decrease in loss values. Remarkably, the employment of CAMuLeNet yielded a substantial enhancement in accuracy, particularly on the English-language CREMA-D benchmark, *with an increase of 10%*, and *an 11% improvement on the RAVDESS benchmark*. Results on CREMA-D are worth noting due to the presence of speakers from various ethnicity uttering emotions at different levels of valence. Hence, Figure-3 contrasts the t-SNE visualizations between Whisper-Medium and CAMuLeNet, with the latter exhibiting pronounced inter-class separation and reduced intra-class variance, which underscores the efficacy of our model in delineating clearer classification boundaries.

Furthermore, the results on multilingual benchmarks, such as CaFE, mirrored the performance gains observed in baseline experiments, underscoring the generalizability of our approach. The multi-task training paradigm notably amplified model's performance on gender recognition tasks, achieving over 95% accuracy on all benchmarks. These findings confirm the effectiveness of our multi-task and co-attention strategy to improve the model's performance on unseen speakers from various linguistic backgrounds.

5.3. Ablation Study

Our proposed methods utilise features from the frequency domain and latent representations generated by a Whisper encoder. Table 1 additionally shows our ablation study by removing the multi-task training setup (1) and then additionally removing co-attention to fuse the features (2). We observed that removing multitask training and replacing it with a single-task training setup reduces performance by an average of approximately 4% over CAMuLeNet. Replacing co-attention based feature fusion with normal concatenation reduced performance across all benchmark datasets below the best-performing baseline methods by an average of around 10% over CAMuLeNet. This could be attributed largely to performing concatenation without accounting for the underlying correlations, emphasising the importance of using co-attention based fusion methods.



a) t-SNE for Whisper Med *b) t-SNE for CAMuLeNet*

Figure 3: *t-SNE visualisation of Feature Distribution. (a) Distribution of final extracted features from Whisper-Medium on the entire CREMA-D dataset. (b) Distribution of final extracted features from CAMuLeNet before the final classification layer on the entire CREMA-D dataset.*

6. Limitations and Conclusion

This study exposes a critical limitation in SER for low-resource languages through our novel BhavVani benchmark, where gains were modest and overall performance was subdued, emphasizing the challenges in unseen speaker recognition. Future work will prioritize these languages and investigate alternative fusion mechanisms for robust generalizations. Our contributions include the introduction of BhavVani dataset, comprehensive benchmark of pretrained embeddings across six datasets using a rigorous 10-fold leave-speaker-out cross-validation strategy, and the novel CAMuLeNet architecture, which synergizes frequency domain features with PTM Whisper embeddings through co-attention based feature fusion and multitask training. These efforts spotlight the issue of emotion recognition of unseen speakers for advancing research in this field.

7. Acknowledgments

We acknowledge the contributions of: Adamya, Aditya, Akshat, Arnav, Arush, Divyanshi, Kartikay, Kuleen, Manaswi, Manya, Rishitej, Rushil, Sahaj, Tanish, Tanishq and Yash from IIIT Delhi in annotating the BhavVani dataset. This work is supported by the Infosys Foundation via Infosys Centre for AI (CAI), IIIT Delhi.

8. References

[1] R. W. Picard, *Affective computing*. MIT press, 2000.

[2] J. D. Mayer, R. D. Roberts, and S. G. Barsade, "Human abilities: Emotional intelligence," *Annu. Rev. Psychol.*, vol. 59, pp. 507–536, 2008.

[3] 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.

[4] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech Emotion Recognition Using Spectrogram & Phoneme Embedding," in *Proc. Interspeech 2018*, 2018, pp. 3688–3692.

[5] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE transactions on multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[6] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[7] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Interspeech 2015*, 2015.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[9] 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.

[10] A. Goel, M. Hira, and A. Gupta, "Multilingual prosody transfer: Comparing supervised & transfer learning," in *The Second Tiny Papers Track at ICLR 2024*.

[11] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe, "End-to-end speech recognition: A survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[12] M. Hira, A. Goel, and A. Gupta, "Crossvoice: Crosslingual prosody preserving cascade-s2st using transfer learning," in *The Second Tiny Papers Track at ICLR 2024*.

[13] Y. He, N. Minematsu, and D. Saito, "Multiple acoustic features speech emotion recognition using cross-attention transformer," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[14] S. Chen, X. Xing, W. Zhang, W. Chen, and X. Xu, "Dwformer: Dynamic window transformer for speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[15] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-Attention for Speech Emotion Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2578–2582.

[16] 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)*, 2022, pp. 6907–6911.

[17] Y. Li, T. Zhao, and T. Kawahara, "Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning," in *Proc. Interspeech 2019*, 2019, pp. 2803–2807.

[18] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech Emotion Recognition with Multi-Task Learning," in *Proc. Interspeech 2021*, 2021, pp. 4508–4512.

[19] C. L. Moine, N. Obin, and A. Roebel, "Speaker attentive speech emotion recognition," *arXiv preprint arXiv:2104.07288*, 2021.

[20] N. Antoniou, A. Katsamanis, T. Giannakopoulos, and S. Narayanan, "Designing and evaluating speech emotion recognition systems: A reality check case study with iemocap," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[21] Z. Zhao, Y. Wang, and Y. Wang, "Knowledge-aware bayesian co-attention for multimodal emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[22] H. Zou, Y. Si, C. Chen, D. Rajan, and E. S. Chng, "Speech emotion recognition with co-attention based multi-level acoustic information," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7367–7371.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[24] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[25] 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.

[26] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

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

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

[29] N. Scheidwasser-Clow, M. Kegler, P. Beckmann, and M. Cernak, "Serab: A multi-lingual benchmark for speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7697–7701.

[30] S. Kumar, A. Kulkarni, M. S. Akhtar, and T. Chakraborty, "When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues," *arXiv preprint arXiv:2203.06419*, 2022.

[31] P. Ekman *et al.*, "Basic emotions," *Handbook of cognition and emotion*, vol. 98, no. 45-60, p. 16, 1999.

[32] 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.

[33] 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*, vol. 16, no. 6, pp. 1505–1518, 2022.

[34] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.