Implicit phonetic information modeling for speech emotion recognition

Tilak Purohit[1,2], *Bogdan Vlasenko*[1], *Mathew Magimai.-Doss*[1]

[1] Idiap Research Institute, Martigny, Switzerland
[2] École polytechnique fédérale de Lausanne (EPFL), Switzerland
tilak.purohit@{idiap,epfl}.ch

Abstract

This study investigates the efficacy of utilizing embedding spaces to model phonetic information in emotion utterances for speech emotion recognition. Our approach involves implicit modeling of phone information by deriving phone-based embeddings from networks specifically trained for phone recognition and pre-trained models fine-tuned for phone/character recognition. The results from evaluating our approach on three speech emotion databases, using both intra-corpus and inter-corpus evaluation methods demonstrate the competitive performance of implicit modeling of phonetic information compared to knowledge-based handcrafted features.

Index Terms: speech emotion recognition, phonetic information, self-supervised learning, fine-tuning

1. Introduction

Over the past two decades, different approaches have emerged for speech emotion recognition (SER) [1], driven by the increasing demand for intelligent systems that can recognize and respond to human emotions, thereby enhancing human-machine interaction. Among them, a few prominent approaches are: (a) extraction of hand-crafted features and modeling them at turn-level or utterance-level through estimation of statistical functionals or obtaining bag-of-audio-word representation [2, 3, 4, 5] (b) direct utterances-level modeling via spectral features or raw-waveforms signals was shown using deep-learning techniques like CNN's, LSTMs and TDNNs for SER [6, 7, 8, 9, 10, 11], and (c) adapting pre-trained self-supervised learning (SSL) based neural networks for SER [12, 13, 14].

Despite the emphasis on modeling emotion at turn-level through hand-crafted features or learning features through neural networks, there has been a constant interest in relating phonetic information and emotion. Lee et al. [15] investigated the impact of emotional coloring on five broad phonetic classes (vowel, glide, nasal, stop, and fricative), they did so by training a phonetic-class based HMM system for each emotional state. The emotion label for the utterance was predicted by first force aligning the input sequence and then comparing the likelihood from each emotion model to determine the emotional state maximizing the likelihood. It was found that vowel sounds as the most effective emotional indicator based on classification performance. Vlasenko et al. [16] used a multi-task learning approach, where two sets of HMM-GMM systems were trained to model phonemes based on two emotional states (high and low arousal). For example, the same phoneme /IY/ based upon the emotional state (high or low arousal) was given two different labels. Classification of emotional state was done during the decoding of the input speech sequence, by taking a majority voting approach for the emotion-labeled phoneme. They demonstrated that phonemes are the smallest possible acoustic units that can classify emotional arousal (high or low). In [17], the phonetic information is modeled for SER by mapping the speech signal and/or word transcription into a sequence of phones and then mapping the sequence of phones into an embedding space using Word2vec [18]. Thereby, explicitly modeling phonetic information through linguistic knowledge-driven embedding space. Dhamyal et al. [19] conducted a systematic study to understand the phonetic composition of emotions. Using a self-attention-based emotion classification model they discovered the most 'attended' phonemes for each emotion class. They reported that the distribution of 'attended'-phonemes tend to vary significantly across natural vs acted emotions. A recent study [20] followed the approach similar to [16, 21] but instead used SSL-based models, which enabled training an emotion-dependent model using a small amount of data. The study investigated the best phonetic units for emotion recognition and showed phonetic units to be helpful and should be incorporated in SER. To explicitly model phonetic information, one would need access to the transcripts or a robust phoneme recognition system. Obtaining speech transcripts incurs overhead, such as the use of human transcribers or the use of speech recognition systems. Furthermore, access to transcripts could raise privacy issues, such as, in speech-based mental health assessment applications.

In this paper, we investigate an alternate approach where phonetic information is implicitly modeled for SER. More precisely, by taking inspiration from recent work on modeling phonetic embeddings for continuous SER in the context of MuSe 2022 challenge [22], we develop a framework where phonetic embeddings are obtained from neural networks pre-tained to classify phonetic or graphemic units and their turn-level statistics are modeled for SER. We study this framework in comparison to the standard approach of modeling hand-crafted features to investigate its potential for speech emotion recognition.

Section 2 presents the study design. Section 3 presents the experimental setup and results. Section 4 presents an analysis of the approach. Finally, Section 5 concludes the paper.

2. Study design

2.1. Methodology

Figure 1 illustrates the approach for implicit modeling of phonetic information in two different manners,

1. extracting phonetic embeddings from neural networks specifically trained to classify phones.
2. extracting phonetic embeddings[1] from SSL networks adapted

[1]For the sake of simplicity, we also refer to graphemic/character embeddings as phonetic embeddings, as grapheme and phoneme are related in spoken language.

on downstream tasks of phoneme recognition or grapheme recognition.

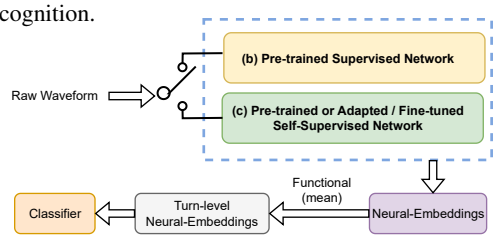

Figure 1: *Proposed neural embedding-based approaches using system (b) and (c). A detailed explanation of (b) can be found in Section 3.1 (b), while (c) is elaborated upon in Section 3.1 (c).*

2.2. Dataset and protocols

To validate the proposed approach, we conduct intra-corpus and inter-corpus speech emotion recognition studies. For that purpose, we employ three standard data sets, namely, EMO-DB, IEMOCAP, and MSP-IMPROV.

Berlin Emotional Speech Database: The Berlin Emotional Speech Database (EMO-DB) [23] covers seven-speaker emotions namely- *anger*, *joy*, *neutral*, *sadness*, *disgust*, *fear*, and *boredom*. The corpus consists of ten (5-male and 5-female) professional actors speaking out ten predefined emotionally neutral sentences. The corpus comprises 900 utterances, where only 493 were marked as 60% natural and 80% assignable by 20 subjects in a listening experiment. We in our study use this subset as suggested in [24].

The Interactive Emotion Dyadic Motion Capture: The Interactive Emotion Dyadic Motion Capture (IEMOCAP) [25] consists of ten (5-male and 5-female) actors over five dyadic sessions performing improvised and scripted scenarios to elicit emotional expressions. To be in line with previous studies we resorted to four emotion classes namely - *angry*, *happy*, *neutral*, and *sad* where we merged the samples from class *excited* with *happy*.

MSP-Improv Database: The MSP-IMPROV [26] corpus comprises recordings from six spontaneous dyadic sessions enacted by twelve actors (6-male and 6-female) from the University of Texas at Dallas. The database claims to carry more naturalness in the recordings. To adhere to previous works we make use of four emotions namely - *angry*, *happy*, *neutral*, and *sad*.

Table 1 summarizes these data sets. To be consistent, for each corpus, we use the protocols that have been used in the literature. More precisely, For EMO-DB, we follow the leave-one-Speaker-out approach. Whereas, for the other two corpora, we use the leave-one-Session-out methodology. That is, for testing the 'k'-th speaker/session, we trained the model on the remaining speakers/sessions. We evaluate the performance of the SER systems in terms of Unweighted average recall (UAR) and Weighted average recall (WAR).

Table 1: *Number of utterances corresponding to each label.*

Database	Content	Ang	Hap	Neu	Sad	Dis	Fea	Bor	Total
EMO-DB	German	127	64	78	52	38	55	79	**493**
IEMOCAP	English	1103	1636	1708	1084	-	-	-	**5531**
MSP-IMPROV	English	792	2644	3477	885	-	-	-	**7798**

3. Experimental setup and results

3.1. System Description

Below, we provide an overview of the systems and their different configurations incorporated for deriving feature representation or neural embeddings by modeling acoustic signals.

(a) Handcrafted feature representation: For the knowledge-based handcrafted feature representation we use COMPARE features [27]. We make use of two configurations of COMPARE features: COMPARE$_{LLD}$ - $65 + 65 = 130$ low–level descriptor($LLDs$) and their delta functions for the frame-level representation and COMPARE$_{LLD\times F}$ - 6373 static turn-level features resulting from the computation of functionals (statistics) over LLD contours. Further, we use the Bag-of-Audio-Words (BOAW) approach implemented in the OPENXBOW toolkit [28] to fetch turn-level representations from COMPARE$_{LLD}$ frame-level representation. In BOAW approach $500 + 500 = 1000$ codebook vectors were created, 500 for 65 LLDs and 500 for 65 LLDs' delta coefficients. This system is denoted as BOAW(COMPARE$_{LLD}$).

(b) Supervised learning based representation: We utilize an off-the-shelf end-to-end Deep CNN based network for phoneme-classification. The network was trained on the 70h of AMI (Augmented Multi-party Interaction) Meeting corpus [29] containing 100h of recordings. The input to the network is a 250ms raw audio signal with a 10ms shift. The network consists of 10 convolutional layers with ReLu activation followed by a fully-connected layer with 1024 neurons, and an output unit with softmax activation for predicting phoneme posteriors. The network was trained for predicting phones based on the triphone modeling, the training was based on cross-entropy loss with stochastic gradient descent and a decaying learning rate. The model provides neural embedding of dimension 1024 corresponding to each 250ms frame. This system is denoted as RAW-CNN(AMI).

(c) Self-supervised learning based representation: We use two different pre-trained self-supervised representation models - Wav2vec2.0 [30] and WavLM [13]. The Wav2vec2.0 model is based on a contrastive model approach, the framework combines contrastive learning with masking. Whereas, WavLM follows a predictive model approach, jointly learning masked speech prediction and denoising in pretraining. For this study, we resort to the base variant of both the models consisting of 12 transformer encoder layers, 768-dimensional hidden states and 8 attention heads comprising 95M and 94.7M parameters for Wav2Vec2.0 and WavLM respectively. Both these models were pre-trained with 960 hours of audio from Librispeech corpus [31]. We utilize these base models to extract the last hidden-state representations with three different settings : (1) The self-supervised pre-trained models (SSPMs) denoted by WAV2VEC2 and WAVLM. (2) SSPMs fine-tuned on TIMIT database [32] for phoneme prediction, denoted by 'SSPM'-FT(TIMIT), and (3) SSPMs fine-tuned on 100h of Librispeech for character classification, denoted by 'SSPM'-FT(LIBRI). S3prl benchmark framework [12] was adopted for setting (1) and (2), and for setting (3) finetuned models were retrieved from HuggingFace[2].

We use support vector machine (SVM) and random forest (RF) as classifiers. To obtain the best baseline systems, the classifiers for handcrafted features underwent hyperparameter tuning using the grid search methodology. For the neural embedding, the parameters were kept the same, i.e., SVM with a linear kernel and RF with gini criterion, so as to ensure a fair comparison among different embedding spaces.

[2]Wav2vec2-base-100h: https://huggingface.co/facebook/wav2vec2-base-100h
WavLM-base-100h: https://huggingface.co/patrickvonplaten/wavlm-libri-clean-100h-base

Table 2: *Comparison of different feature representations for emotion recognition on three evaluation corpora. Group-1(G-1): Knowledge-based handcrafted features. Group-2(G-2): Supervised learning (SL) based features. Group-3.1(G-3.1): Self-supervised learning (SSL) Wav2vec2 based features. Group-3.2(G-3.2): SSL WavLM based features.*

			Evaluation corpus					
			IEMOCAP (4-class)		MSP-IMPROV (4-class)		EMO-DB (7-class)	
Feature representation	**Dim.**	**Classifier**	**UAR**	**WAR**	**UAR**	**WAR**	**UAR**	**WAR**
Group -1								
ComParE$_{LLD\times F}$	6373	SVM	58.00	56.51	43.10	55.90	80.20	81.91
ComParE$_{LLD\times F}$	6373	RF	58.55	57.43	36.21	55.69	66.31	72.97
BoAW(ComParE$_{LLD}$)	500/500	SVM	57.67	56.62	43.30	55.60	70.85	73.44
BoAW(ComParE$_{LLD}$)	500/500	RF	58.36	57.41	35.87	55.27	57.19	64.52
Group -2								
Raw-CNN(ami)	1024	SVM	59.10	58.22	44.33	59.20	77.48	79.72
Raw-CNN(ami)	1024	RF	52.18	51.89	37.19	56.49	69.16	73.02
Group -3.1								
Wav2vec2	768	SVM	62.09	61.38	48.60	59.84	83.71	85.40
Wav2vec2	768	RF	53.27	52.63	38.94	57.19	65.73	72.62
Wav2vec2-ft(timit)	768	SVM	57.68	56.34	45.89	58.69	67.35	70.79
Wav2vec2-ft(timit)	768	RF	48.47	48.13	34.25	58.69	49.35	57.81
Wav2vec2-ft(libri)	768	SVM	62.46	61.25	51.14	61.96	75.24	76.27
Wav2vec2-ft(libri)	768	RF	51.97	51.00	37.66	53.60	59.03	64.91
Group -3.2								
WavLM	768	SVM	**64.38**	**63.41**	**54.40**	**64.64**	**87.67**	**88.44**
WavLM	768	RF	56.99	56.73	38.99	57.94	68.51	74.44
WavLM-ft(timit)	768	SVM	57.44	56.73	45.69	58.44	63.12	66.33
WavLM-ft(timit)	768	RF	47.12	46.94	34.26	52.54	44.69	52.33
WavLM-ft(libri)	768	SVM	60.73	59.61	49.19	61.36	60.22	63.69
WavLM-ft(libri)	768	RF	48.37	47.89	36.12	52.42	42.59	46.86
Early fusion for selected systems								
G-1+G-2	6373 + 1024	SVM	59.71	58.16	47.63	57.84	80.00	83.40
G-1+G-3.1 (libri)	6373 + 768	SVM	60.62	59.15	50.54	59.72	82.39	84.02
G-1+G-3.2 (libri)	6373 + 768	SVM	60.79	59.29	49.98	59.24	84.25	85.68

3.2. System Performance

Table 2 presents the performance of different systems for the intra-corpus study. Embeddings generated via Wav2vec2 and WavLM outperform other systems for all three corpora, with WavLM delivering the best performance for the SER task. When considering phonetic embeddings, it is interesting to observe that the embeddings derived from Raw-CNN(ami) network and SSL networks adapted for phoneme/character recognition, 'SSPM'-ft(timit) and 'SSPM'-ft(libri) consistently outperform the knowledge-based handcrafted features on IEMOCAP and MSP-IMPROV, in particular SVM classifier. In the case of German-speaking corpus EMO-DB, however, we do not observe this trend. This suggests that the phonetic embeddings do not generalize well in cross-lingual scenarios. Having said that, when the hand-crafted features and phonetic embeddings are combined through early fusion, we observe that the performance of systems remains steady. In the case of MSP-IMPROV, we observe considerable gains when fusing Group-1 (G-1) and Group-2 (G-2) features, when compared to the standalone G-1 and G-2 feature representations. Finally, we can observe that SVM is able to better model the phonetic embeddings than RF for SER. One possible reason for that could be that RFs partition the feature space using a series of decision trees, which may not be effectively capturing the non-linear relationships between the features in the embedding space.

It is worth mentioning, the performance for handcrafted features reported in Table 2 (G-1 features) are comparable to the results previously reported in the literature, for EMO-DB [24, 33], for IEMOCAP [9, 34], and MSP-IMPROV [26, 35]. Similarly, the performance obtained with Wav2vec2 and WavLM embedding is consistent with previous studies reported for IEMOCAP and EMO-DB corpus [12, 13, 14, 36, 37]. Due to space limitations, we are not able to present the confusion-matrix analysis of different systems. This supplementary information can be found on the github page [3]

4. Analysis

4.1. Inter corpus training analysis

For the inter-corpus training evaluation, we use Group-1 features for baseline and selected the best-performing phonetic information-based system from each of Group-2 and 3 in Table 2. We also conducted early-fusion experiments. The inter-corpus experiments were carried out among two databases IEMOCAP and MSP-IMPROV since they both have the same emotion classes (*angry*, *happy*, *neutral*, and *sad*) and they are based on dyadic conversation. SVM was chosen as the classifier for this analysis.

The relatively lower performance reported in Table 3 for inter-corpus training compared to the intra-corpus training performance as in Table 2 highlights the challenge associated with generalizing emotion across cross-domain databases. Nevertheless, we can observe that phonetic embedding obtained from the SSL network generalizes across the two corpora better than hand-crafted features. Early fusion of these two features yields a stable performance despite standalone features yielding inferior performance. This indicates that in early fusion the classifier is giving more emphasis to phonetic embeddings than hand-crafted features.

4.2. Impact of ASR accuracy

From Table 2, it can be seen that the SSL-based embedding performance for SER task decreases after we fine-tune the systems

[3] Supplementary material: gitlab.idiap.ch/emil/is-2023_ph-ser

Table 3: *Performance comparison of different feature representation for the 4-class classification task in inter-corpus training scheme.*

Systems	Dim.	UAR	WAR
Train IEMOCAP Test MSP-IMPROV			
COMPARE$_{LLD\times F}$	6373	38.83	36.86
BOAW(COMPARE$_{LLD}$)	500/500	41.32	39.71
RAW-CNN(AMI)	1024	32.19	52.13
WAV2VEC2-FT(LIBRI)	768	37.69	45.70
WAVLM-FT(LIBRI)	768	44.70	53.53
G1+G2	6373 + 1024	43.31	40.14
G1+G3.1.(LIBRI)	6373 + 768	**46.97**	**48.21**
G1+G3.2.(LIBRI)	6373 + 768	41.37	45.20
Train MSP-IMPROV Test IEMOCAP			
COMPARE$_{LLD\times F}$	6373	41.41	43.33
BOAW(COMPARE$_{LLD}$)	500/500	38.16	40.31
RAW-CNN(AMI)	1024	32.98	35.80
WAV2VEC2-FT(LIBRI)	768	**46.85**	**48.82**
WAVLM-FT(LIBRI)	768	44.66	48.44
G1+G2	6373 + 1024	35.47	38.02
G1+G3.1.(LIBRI)	6373 + 768	45.34	46.71
G1+G3.2.(LIBRI)	6373 + 768	44.58	46.64

with TIMIT for phoneme recognition ('SSPM'-FT(TIMIT)). However, it improves (compared to 'SSPM'-FT(TIMIT)) when the SSPMs are fine-tuned on the 100-h Librispeech corpus ('SSPM'-FT(LIBRI)) for character recognition, with the exception of WAVLM for EMO-DB. These initial results suggest that having more data with greater speaker variability may help improve performance for the cross-domain SER task. To further investigate this, we conducted an independent experiment, previously we observed that in the case of WAV2VEC2-FT(LIBRI) the fine-tuning on the larger corpus (100-h Librispeech) helped attain performance comparable to WAV2VEC2. Therefore, we decided to use the same SSL variant (Wav2vec2.0 base) but fine-tuned on an even larger set, 960-h Librispeech corpus. As expected, this model provided better results for the in-domain task (character recognition) with a lower word error rate of 3.4% on Librispeech 'clean' set in comparison to 6.1% based on 100-h tuning. We then evaluated the extracted embeddings for the SER task. We found that the performance degrades, with UARs of 48.89, 40.18, and 54.54 for IEMOCAP, MSP-IMPROV, and EMO-DB, respectively, when compared to UARs of 62.46, 51.14, and 75.24 (Table 2) 100hr fine-tuned net. This indicates that arbitrarily increasing the data does not necessarily yield phonetic embeddings informative for SER. As a by-product, this analysis shows that fine-tuning the SSL model on a large amount of data for speech recognition is making the embedding space less invariant to emotional differences.

4.3. Embedding space analysis

While Wav2vec2.0 was originally meant for the ASR task, WavLM was positioned as a full-stack speech processing model [13]. Table 2 shows embeddings generated via WAV2VEC2 and WAVLM outperform other systems, this is not surprising, given that these SSL representations are well recognized for carrying rich speech information and have demonstrated strong generalization and competitiveness across various downstream tasks [12].

Despite the superior performance of WAVLM based embedding as seen in Table 2, it is noteworthy that it exhibits a greater loss of emotional content compared to WAV2VEC2-FT(LIBRI) when fine-tuned for character recognition using 100-h Librispeech. This is evident from the relatively lower performance of WAVLM-FT(LIBRI) compared to WAV2VEC2-FT(LIBRI). This behavior can be attributed to the fact that after fine-tuning, WAVLM emphasizes more on spoken content modeling and speaker identity preservation [13], while discarding the paralinguistic content that carries emotional information. Consequently, the embedding space becomes less expressive of emotions, which can explain the drop in performance for the cross-domain SER tasks.

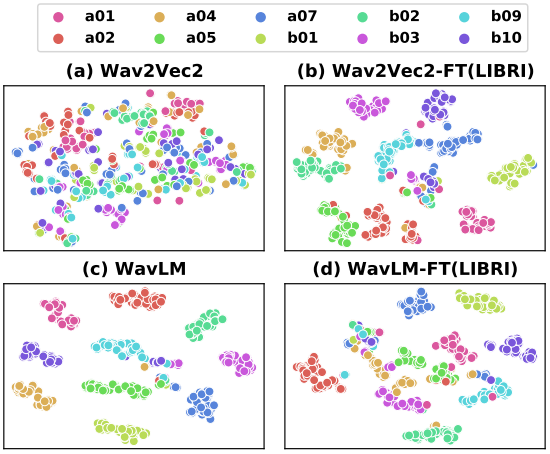

Figure 2: *t-SNE plots for different embeddings spaces before and after finetuning for the phoneme recognition task. Labels aXX and bXX correspond to text IDs' used in the* EMO-DB *database.*

One can observe from Table 2 that the embedding produced by RAW-CNN(AMI) outperforms SSL-based phonetic embedding in the case of EMO-DB. This observation could be explained by the fact that RAW-CNN(AMI) is trained to predict context-dependent tri-phones, which appears to make the embedding space more robust.

Finally, to visualize the embedding space we use embeddings generated via SSPM's before and after finetuning for the phoneme recognition task. We generate embedding for EMO-DB, since it is a phonetically balanced database. For the case of WAVLM, we observe clusters corresponding to sentence ID's whereas this is not seen for the case of WAV2VEC2. Once the models are finetuned for phoneme recognition these clusters can be seen for both the cases. This suggests that WAVLM is modeling spoken content and speaker identity [13] without any fine-tuning, unlike WAV2VEC2 which may explain WAVLM performance in Table 2.

5. Conclusions

In this paper, we proposed an approach to implicitly model phonetic information for speech emotion recognition. We do so by utilizing networks specifically trained for the phoneme recognition task and pre-trained models fine-tuned for the phoneme/character recognition task to extract phonetic-embeddings; model their statistics at turn level; and classify them using SVM or RF. Our experimental studies on three corpora show that phonetic embeddings can yield competitive SER systems when compared to hand-crafted features, and these two features could be jointly modeled for robust SER. Our analysis of the phonetic embedding space indicated that there is an inverse relation between speech recognition performance and the effectiveness of the extracted phonetic embeddings for SER. Our future work will investigate this aspect further to understand what kind and amount of phonetic units classification data best suits SER.

6. Acknowledgements

This work was funded by the SNSF through the Bridge Discovery project EMIL: Emotion in the loop - a step towards a comprehensive closed-loop deep brain stimulation in Parkinson's disease (grant no. $40B2-0\_194794$).

7. References

[1] B. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, 2018.

[2] B. Schuller, A. Batliner, D. Seppi *et al.*, "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals," in *Proc. of Interspeech*, 2007.

[3] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International journal of speech technology*, 2012.

[4] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," *International Journal of Speech Technology*, 2018.

[5] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, 2020.

[6] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation Learning for Speech Emotion Recognition," in *Proc. of Interspeech*, 2016.

[7] M. Neumann and N. T. Vu, "Attentive Convolutional Neural Network Based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech," in *Proc. of Interspeech*, 2017.

[8] Z. Peng, Y. Lu, S. Pan, and Y. Liu, "Efficient Speech Emotion Recognition Using Multi-Scale CNN and Attention," in *Proc. of ICASSP*, 2021.

[9] T. Purohit, S. Yadav, B. Vlasenko, S. P. Dubagunta, and M. Magimai.-Doss, "Towards Learning Emotion Information from Short Segments of Speech," in *Proc. of ICASSP*, 2023.

[10] J.-L. Li, T.-Y. Huang, C.-M. Chang, and C.-C. Lee, "A Waveform-Feature Dual Branch Acoustic Embedding Network for Emotion Recognition," *Frontiers in Computer Science*, 2020.

[11] P. Kumawat and A. Routray, "Applying TDNN Architectures for Analyzing Duration Dependencies on Speech Emotion Recognition," in *Proc. of Interspeech*, 2021.

[12] 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," in *Proc. of Interspeech*, 2021.

[13] 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*, 2022.

[14] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," *Proc. Interspeech 2021*, 2021.

[15] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes," in *Proc. of Interspeech*, 2004.

[16] B. Vlasenko and A. Wendemuth, "Determining the smallest emotional unit for level of arousal classification," in *Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013.

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

[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, ICLR*, 2013.

[19] H. Dhamyal, S. A. Memon, B. Raj, and R. Singh, "The phonetic bases of vocal expressed emotion: natural versus acted," in *Proc. of Interspeech*, 2020.

[20] J. Yuan, X. Cai, R. Zheng, L. Huang, and K. Church, "The role of phonetic units in speech emotion recognition," *arXiv preprint arXiv:2108.01132*, 2021.

[21] B. Vlasenko, D. Prylipko, R. Böck, and A. Wendemuth, "Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications," *Computer Speech & Language*, 2014.

[22] S. Yadav, T. Purohit, Z. Mostaani, B. Vlasenko, and M. Magimai.-Doss, "Comparing Biosignal and Acoustic feature Representation for Continuous Emotion Recognition," in *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, 2022.

[23] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. of Interspeech*, 2005.

[24] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Proc. of ASRU*, 2009.

[25] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, 2008.

[26] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, 2017.

[27] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. of Interspeech*, 2013.

[28] M. Schmitt and B. Schuller, "Openxbow: introducing the Passau open-source crossmodal bag-of-words toolkit," *Journal of Machine Learning Research*, 2017.

[29] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," *Language Resources and Evaluation*, 2007.

[30] 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*, 2020.

[31] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. of ICASSP*, 2015.

[32] J. Garofolo *et al.*, "TIMIT Acoustic-phonetic Continuous Speech Corpus," *Linguistic Data Consortium*, 1993.

[33] F. Eyben, K. R. Scherer, B. W. Schuller *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, 2016.

[34] S. Amiriparian, A. Sokolov, I. Aslan, L. Christ, M. Gerczuk *et al.*, "On the impact of word error rate on acoustic-linguistic speech emotion recognition: An update for the deep learning era," *arXiv:2104.10121*, 2021.

[35] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *Proc. of ICASSP*, 2019.

[36] A. Keesing, Y. S. Koh, and M. Witbrock, "Acoustic features and neural representations for categorical emotion recognition from speech." in *Proc. of Interspeech*, 2021.

[37] O. C. Phukan, A. B. Buduru, and R. Sharma, "A comparative study of pre-trained speech and audio embeddings for speech emotion recognition," *arXiv:2304.11472*, 2023.