COMPARING DATA-DRIVEN AND HANDCRAFTED FEATURES FOR DIMENSIONAL EMOTION RECOGNITION

Bogdan Vlasenko *Sargam Vyas* *Mathew Magimai.-Doss*

Idiap Research Institute, Martigny, Switzerland

ABSTRACT

Speech Emotion Recognition (SER) has garnered significant attention over the past two decades. In the early stages of SER technology, 'brute force'-based techniques led to a significant expansion in knowledge-based acoustic feature representation (FR) for modeling sparse emotional data. However, as deep learning techniques have become more powerful, their direct application has been limited by the scarcity of well-annotated emotional data. As a result, pre-trained neural embeddings on large speech corpora have gained popularity for SER tasks. These embeddings leverage existing transfer learning methods suitable for general-purpose self-supervised learning (SSL) representations. Recent studies on downstream SSL techniques for dimensional SER have shown promising results. In this research, we aim to evaluate the emotion-discriminative characteristics of neural embeddings in general cases (out-of-domain) and when fine-tuned for SER (in-domain). Given that most SSL techniques are pre-trained primarily on English speech, we plan to use speech emotion corpora in both language-matched and mismatched conditions. We will assess the discriminative characteristics of both handcrafted and standalone neural embeddings as FRs.

Index Terms— Self-supervised learning (SSL), emotion recognition, dimensional emotion, VAD

1. INTRODUCTION

Speech-based emotion recognition (SER) is a key technology for facilitating user-centered human-to-machine interactions [1]. SER research is dominated by two conceptual emotion representation techniques: discrete emotions and emotional dimensions. The first investigates discrete emotional categories (anger, joy, etc.), while the latter focuses on the dimensions of valence, arousal, and dominance (VAD). In the case of VAD based SER modelling emotional state could be represented with continuous scalar values for valence, arousal and dominance emotional scale. The developments of dimensional and discrete turn-level SER techniques have had different tendencies and used different classification measuring techniques. In neuropsychological science, the neural processes that correspond to linguistic and acoustic (verbal) information remain undiscovered [2]. Empirical studies on dimensional SER show that combining acoustic and linguistic modeling could improve the performance of emotional valence modeling [3]. The evolution and development of discrete classes of emotion recognition were supported by the advancement of the Computational Paralinguistic Challenge (ComParE) introduced in 2009 [4]. To support interest in ComParE challenges, a free toolkit for modeling turn-level features was introduced: OpenSMILE [5]. Most of the ComParE challenges used turn-level ground truth labels. Hence, a list of indicative turn-level feature representations was redefined during the first ComParE challenges. The "brute force" concept was quite a popular technique used for selecting the most informative acoustic features for each challenge. The saturation point for hand-crafted turn-level feature representations was reached in 2013 with the introduction of the ComParE 2013 [6] feature set, which comprises 6,373 features.

In the case of dimension emotional modeling, research studies were oriented not only on selecting appropriate signal processing techniques but also on emotion perception and annotation. By using dimensional annotation techniques affective computing community was targeted modeling spontaneous naturalistic emotions annotated by a pool of annotators. One decade ago, in earlier studies on dimensional SER [7], the authors showed that a subset of the emotional corpora with more reliable annotations could provide better dimensional emotion modeling. Even considering the long history of acoustic-based SER there are several open research questions on acoustic emotional theory. The smallest acoustic emotional unit has not been defined. In the case of discrete SER one could make the assumption that emotional content is equally distributed between phonemes and apply explicit phoneme-level emotion modeling [8–10]. In the case of explicit phoneme-level emotion modeling emotional models were trained on sub-word level units for each emotional class. For discrete emotion modeling phoneme-level emotion modeling provides quite a comparative performance.

In recent years, the artificial intelligence (AI) field is undergoing a major paradigm shift, moving from task-specific architectures trained for a given case to general-purpose foundation models that can be applied to several use cases [11]. Taking into account the classification performance of discrete SER presented in various benchmark studies [12–15] we noticed that out-of-domain neural embedding could provide important information for SER. Implementation of downstream and fine-tuning of self-supervised learning techniques become a new trend in discrete and dimensional SER. In [16] authors mentioned that a new era in dimensional SER is starting with pre-trained, transformer-based foundation models, which could encapsulate dominant information streams of spoken language, linguistics, and paralinguistics. Considering the previous study [17] on combining SSL with handcrafted features we would like to check if these two types of FR could improve performance for dimensional SER. In contrast to the evaluation of different transformer layers for SER [17, 18] we used just a final layer outputs for acoustic emotion modeling.

Considering recent developments in uncertainty modeling for dimensional SER [19] we assume that the number of available emotional corpora with reliable dimensional emotion annotations is comparable low. The first corpora that we select for our study is the VAM ("Vera am Mittag") [20] corpus which uses 17 and 6 annotators for dimensional labeling and uses evaluator weighted estimator (EWE) [21] for smoothing uncertainty effect in multi-annotators setup. Presented earlier study [22] showed that the spectral content of syllabic nuclei could provide a promising SER emotion recognition for the VAM. The second dataset that we are using, the IEMO-

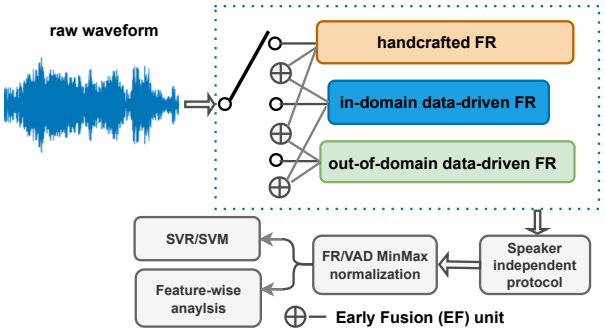
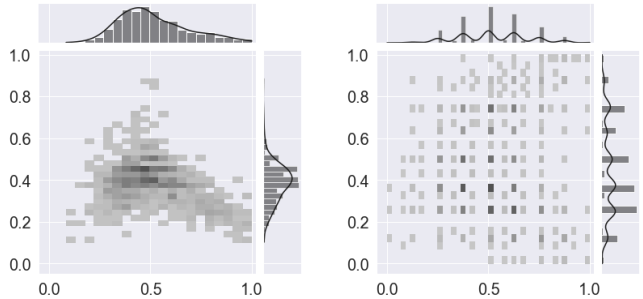


Fig. 1. Processing flow of proposed experimental study.



(a) VAM (b) IEMOCAP

Fig. 2. Data distribution: arousal (horizontal), valence (vertical)

CAP (Interactive Emotional Dyadic Motion Capture) contains aggregated VAD labels provided by comparable lower number of annotators. Hence, an uncertainty modeling for dimensional SER on the IEMOCAP and MSP-Podcast has been addressed in [19]. In [23] the authors mentioned that more focus should be placed on ordinal regression instead of classifying or predicting discrete emotional states within SER tasks, which are the common practices to date. More attention should be addressed to the quality and reliability of dimensional emotional annotations for naturalistic emotional corpora with turn-level annotations. We compare obtained results with state-of-the-art [3, 24–28] and discuss uncertainty issues in dimensional SER.

2. METHODS

This section describes the acoustic FR derived from audio signals, machine learning architecture, and measures used for dimensional SER. The processing flow of our dimensional SER is presented in Figure 1

Features representation: For the knowledge-based handcrafted FR, we use CMP - COMPARE 2016 [6]. Feature set contains 6373 static turn-level features resulting from the computation of functionals (statistics) over Low-Level Descriptors contours. Considering top performance positions on challenge leader-boards for out-of-domain SSL embeddings on SER task for SUPERB challenge [12] we employed a wide range of SSL embeddings: WAV2VEC2 Large Robust (WV2 LR) [29], HUBERT (large) (HRT) [30], WAVLM (large) (WLM) [31]. Finally, for in-domain SSL modeling, we used WAV2VEC2 (WV2 EM) fine-tuned for dimensional SER on MSP-Podcast dataset [32]. A fixed-length utterance-level feature representation is obtained by computing mean of the frame-level SSL embeddings extracted from the final layer. We applied the Early Fusion (EF) technique to see if a combination of top-performing FR could improve discriminative characteristics.

Regressors: The SVM-based regression technique for the evaluation of discriminative characteristics of selected FR. Support Vector Regression (SVR) for the regression task of predicting VAD levels was used. The radial basis function (RBF) kernel and MinMax for feature normalization were applied.

3. EXPERIMENTAL SETUP

The section introduces the multi-lingual emotional corpora, experiential protocols, and evaluation metrics used for the study.

Corpora: The VAM corpus [20] contains 947 emotional speech samples collected from 47 German speakers (11m/36f). Speech samples were selected from 12 broadcasts of the talk show "Vera am Mittag" (in English, "Vera at noon"). The weighted average values with EWE techniques of valence, arousal, and dominance emotional dimensions were used as the ground-truth labels of each sentence. In addition, we used the English dataset - IEMOCAP [33]. The corpus includes 10,039 emotional speech samples collected from recordings captured during 5 dyadic interaction sessions. Ground-truth values were determined by taking the average of all annotators who participated in the emotion perception study. Figure 2 represents a distribution of aggregated valence and arousal samples in selected emotional corpora. A comparably small pool of annotators used for dimensional labeling of IEMOCAP resulted in a high level of granularity of aggregated labels. Also, a major part of emotional samples for VAM are located in the low valence area.

Metrics: For evaluating SER performance we used: Concordance correlation coefficient (CCC), (see eq. 1) and Root mean square error (RMSE).

$$CCC(X,Y) = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (1)$$

where $\rho = \frac{COV(x,y)}{\sigma_x \cdot \sigma_y}$ is a Pearson correlation coefficient, and μ and σ are the mean and standard deviation, respectively. Argument X denotes the ground-truth VAD labels, and argument Y denotes either the predicted emotional dimension's level (refer to the results in Table 1) or the min-max normalized FR observation (see section 5). For both datasets, original aggregated VAD labels were mapped into the interval 0 to 1.

Evaluation protocol: Similar to the previous studies on selected corpora, we conducted speaker-independent experiments following the Leave-One-Session/Speaker group-Out (LOSsO/LOSgO) methodology for training regression models. We trained models on speech samples from 4 Sessions and used the remaining 5th session for testing. For both datasets, test predictions obtained during 5 test folds were accumulated and used to estimate overall CCC and RMSE rates.

Implementation tools: The COMPARE FR were extracted with implemented in python opensmile package [5]. In-domain (WAV2VEC) models for down-stream-based [32] direct prediction of VAD rates and denoted obtained results with E2E EM. Pre-trained out-of-domain models were downloaded from huggingface repositories. Supplementary material and Python codes can be found in the following GitHub repository[1].

[1]https://github.com/idiap/ICASSP24_Dim_SER

FR/ Metod	Valence CCC	Valence RMSE	Arousal CCC	Arousal RMSE	Dominance CCC	Dominance RMSE
VAM - *language mismatched condition*						
1.CMP	0.235	0.09	0.774	0.10	0.762	0.09
2.WV2 LR	0.283	0.09	0.773	0.11	0.769	0.09
3.WV2 EM	**0.324**	0.09	0.768	0.10	0.759	0.09
4.HRT	0.196	0.10	0.700	0.11	0.669	0.10
5.WLM	0.211	0.10	0.688	0.12	0.680	0.10
EF(1.+2.)	0.261	0.09	0.781	0.10	0.771	0.09
EF(1.+3.)	0.317	0.09	**0.805**	0.10	**0.800**	0.09
EF(1.+5.)	0.252	0.09	0.784	0.10	0.775	0.09
E2E EM	-0.121	0.16	0.260	0.35	0.284	0.28
IEMOCAP - *language matched condition*						
1.CMP	0.379	0.20	0.667	0.13	0.488	0.17
2.WV2 LR	0.424	0.19	0.686	0.12	0.519	0.16
3.WV2 EM	**0.683**	0.16	0.702	0.12	0.531	0.16
4.HRT	0.584	0.17	0.694	0.12	0.527	0.17
5.WLM	0.603	0.17	0.701	0.12	**0.535**	0.16
EF(1.+2.)	0.403	0.20	0.676	0.13	0.502	0.17
EF(1.+3.)	0.610	0.17	0.693	0.12	0.520	0.17
EF(1.+5.)	0.502	0.19	0.687	0.12	0.513	0.17
EF(3.+5.)	**0.683**	0.16	**0.708**	0.12	**0.535**	0.16
EF(3.+4.)	0.680	0.16	**0.707**	0.12	**0.535**	0.16
EF(4.+5.)	0.609	0.17	0.704	0.12	0.533	0.16
E2E EM	0.478	0.21	0.660	0.147	0.486	0.18

Table 1. CCC and RMSE rates for the VAM and IEMOCAP studies. Abbreviations: EF - early fusion

4. RESULTS

During the first experimental phase, we used the VAM dataset for simulating language mismatched condition. Direct predictions obtained with E2E emotional models trained on English speech could not provide applicable emotion regression performance for German emotional speech samples. On the other hand, in-domain WAV2VEC2 EM could boost regression performance for the valence emotional dimension. The WAV2VEC2 LR which represents the out-of-domain FR, provides competitive CCC rates for arousal and dominance emotional dimensions.

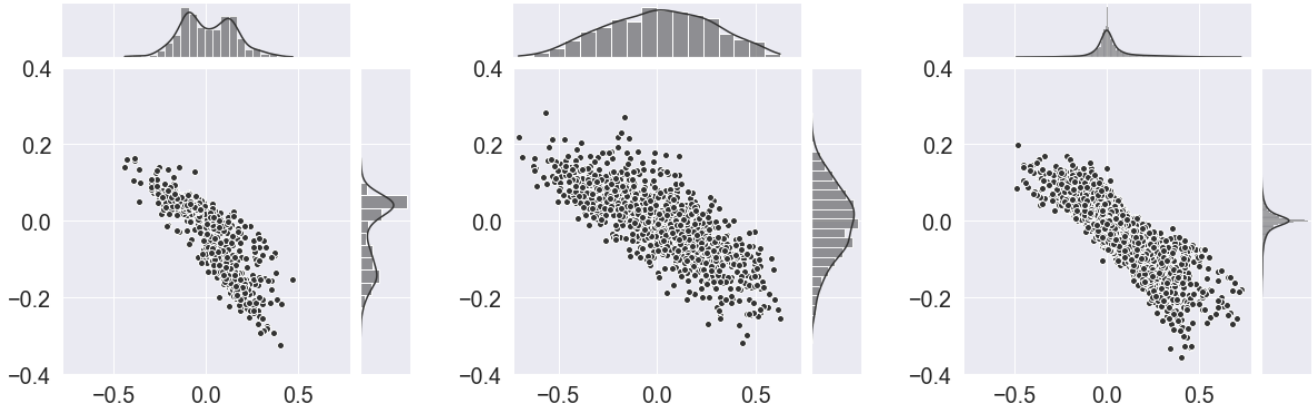
The highest average CCC (overall VAD dimensions) were observed for WAV2VEC2 EM, WAV2VEC2 LR, COMPARE and WAVLM. Considering different types of top-performing FR we decided to do an early fusion study by combining handcrafted and data-driven features. As one could see from Table 1, a combination of knowledge-based and in-domain WAV2VEC2 EM FR provide the highest CCC rates for arousal and dominance dimensions. For the valence emotional dimension, the best CCC was obtained with raw WAV2VEC2 EM FR. In [22] authors used feature selection and parameters tuning for 10-fold cross-validation. In our study, we used speaker-independent protocol and avoided hyper-parameters tuning. We used CCC rates instead of Pearson correlation for measuring emotion classification performance. In our study, we used a larger set of FR and obtained comparable performance.

During the second experimental phase, we simulated matched language condition and conducted a study on the IEMOCAP dataset. As one can see from Table 1, E2E direct modeling provides applicable emotion regression performance, comparable with performance reported in [16]. Still, in-domain FR provides significantly better CCC rates for valence dimension in comparison with E2E approach. The highest average CCC (overall VAD dimensions) were observed for data-driven FR: out-of-domain HUBERT, WAVLM, and in-domain WAV2VEC2 EM. Early fusion of in-domain and out-of-domain FRs improves CCC rates for all emotional dimensions. Presented results are comparable with state-of-the-art-results presented in [3, 24, 28]. Even advanced uncertainty modeling for dimensional SER presented in [19] reported lower CCC rate for valence prediction (CCC=0.625 on 5 folds).

5. FEATURE REPRESENTATION ANALYSIS

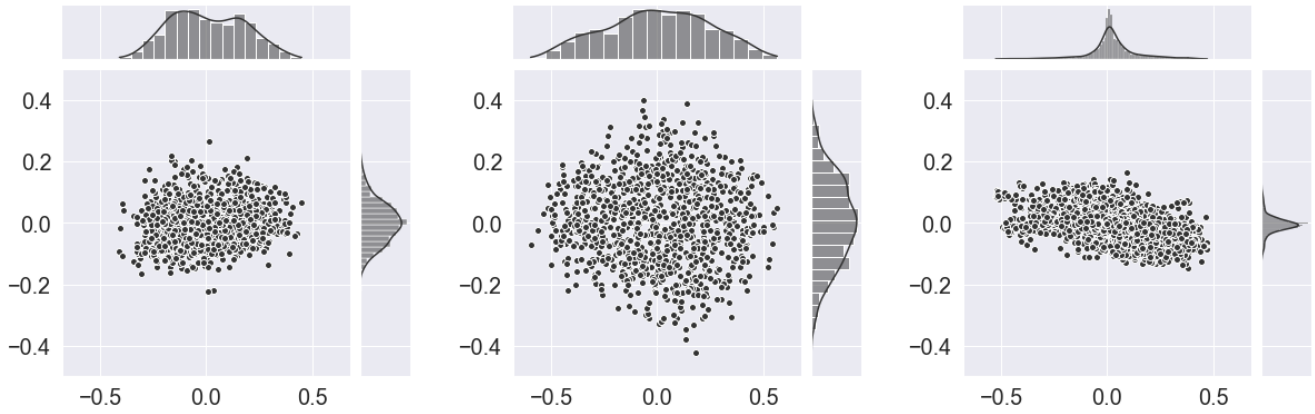
In order to evaluate the discriminative characteristics of employed knowledge-based and data-driven FR we applied MinMax normalization to features and VAD dimensional labels. Afterward, we estimated CCC rates for each FR.

Figure 3 shows that out-domain data-driven FRs pre-trained on English emotional speech provide discriminative information for German emotion arousal modelling. Fine-tuned WAV2VEC2 EM provides additional emotion-related information for arousal and valence emotional dimensions. Knowledge-based FR provides quite high discriminative characteristics on the VAM speech samples, considering a more reliable concept used for dimensional emotion



(a) wv2 LR (b) wv2 em (c) ComParE

Fig. 3. VAM: distribution of feature-wise CCC rates. Arousal (horizontal axis) and valence (vertical axis).



(a) wavLM (b) wv2 em (c) ComParE

Fig. 4. IEMOCAP: distribution of feature-wise CCC rates. Arousal (horizontal axis) and valence (vertical axis).

annotation.

Results obtained on the IEMOCAP database, see Figure 4, show that out-of-domain SSL representations for WAVLM provide a comparable lower range of CCC rates for valence. And in-domain WAV2VEC2 EM outperforms out-of-domain SSL for valence. On the other hand, EF results presented in Table 1 show that in-domain and out-of-domain SSL are complimentary for arousal and dominance emotional dimensions. At the same time, feature-wise CCC rates for out-of-domain SSL FR have a narrower range in arousal dimension. RMSE errors reported for the VAM studies are significantly lower in comparison with the IEMOCAP studies. On the other hand, the IEMOCAP dataset contains around 10 times more emotional speech samples and should provide better emotion modeling opportunities and lower RMSE rates. As one can see from Table 2 more reliable annotation labels assigned to the VAM dataset provide us better feature-wise CCC rates. Also, top indicative knowledge-based FR have significantly higher feature-wise CCC for VAM then IEMOCAP database. For both datasets, one from the top two discriminative FR for arousal and dominance emotional dimensions are the same: IEMOCAP - selected functional from mfcc_sma[2] LLD and VAM - selected functional to audspec_lengthL1norm LLD. Hence, we assume that arousal and dominance emotional dimensions share some common acoustic characteristics. Our investigations confirm the results presented in [19, 22] that spectral features are the most discriminative for modeling emotional dimensions.

6. CONCLUSION

We investigated different data-driven and handcrafted FR for dimensional SER. To evaluate discriminative characteristics of in-domain and out-of-domain data-driven FR we used German and English emotional speech corpora. The obtained results showed that in-domain FR fine-tuned for modeling English emotional speech could provide additional sources of information handcrafted FR during dimensional SER on German speech. Experimental studies on English emotional speech showed that in-domain and out-of-domain SSL-based FR provides complementary sources of information. Results for direct prediction with in-domain WAV2VEC2 fine-tuned on MSP-Podcast dataset shows a domain difference not only for language-mismatched conditions but also for language-matched experimental setup. Finally, we showed that stand-alone data-driven FR for in- and out-of-domain models could be used for dimensional SER.

Dim.	Feature representation	CCC
	VAM	
Val	mfcc_sma[1]_upleveltime75	0.199
Val	mfcc_sma[3]_percentile1.0	0.171
Aro	audspec_lengthL1norm_sma_quartile3	0.725
Aro	audspec_lengthL1norm_sma_peakMeanAbs	0.716
Dom	audspec_lengthL1norm_sma_percentile99.0	0.659
Dom	audspec_lengthL1norm_sma_peakMeanAbs	0.669
	IEMOCAP	
Val	audspec_lengthL1norm_sma_de_flatness	0.166
Val	audSpec_Rfilt_sma_de[23]_flatness	0.144
Aro	mfcc_sma[2]_range	0.469
Aro	mfcc_sma[2]_pctlrange0-1	0.473
Dom	mfcc_sma[2]_pctlrange0-1	0.375
Dom	mfcc_sma_de[4]_lpc1	0.355

Table 2. Feature-wise CCC rates for top-performing handcrafted FRs.Abbreviation: Dim. - dimensionality.

Obtained RMSE rates for the VAM and IEMOCAP studies show that providing more reliable dimensional emotion labeling could be more beneficial than using more emotional speech samples with less reliable emotional labels. In our internal research project studies we compensate for an uncertainty effect in dimensional emotion annotation by using proper emotion perception tests and employing a large pool of emotion annotators. Qualitative analysis of feature-wise correlation plot rates shows interesting tendencies in in- and out-of-domain data-driven FR for matched and mismatched language setups.

In our future work, we are planning to evaluate in- and out-of-domain stand-alone data-driven FR for continuous dimensional emotion recognition on datasets provided by Multimodal Sentiment Analysis Challenge organizers [34] with more advanced regression and feature transformation techniques [35]. We are planning to utilize the presented FR for emotional cue transfer in Text-To-Speech system modeling.

7. ACKNOWLEDGEMENTS

This work was partially funded by the Swiss National Science Foundation through the Bridge Discovery project EMIL (grant no. $40B2-0\_194794$) and by the Innosuisse through the Flagship project IICT (grant agreement no. $PFFS-21-47$).

8. REFERENCES

[1] Björn W Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[2] Celine Berckmoes and Guy Vingerhoets, "Neural foundations of emotional speech processing," *Current Directions in Psychological Science*, vol. 13, no. 5, pp. 182–185, 2004.

[3] Bagus Tris Atmaja and Masato Akagi, "Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using SVM," *Speech Communication*, vol. 126, pp. 9–21, 2021.

[4] Björn Schuller et al., "The INTERSPEECH 2009 emotion challenge," in *Proc. Interspeech 2009*, 2009, pp. 312–315.

[5] Florian Eyben et al., "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proc. ACMM MM*, 2010, pp. 1459–1462.

[6] Björn Schuller et al., "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. Interspeech*, 2013.

[7] Siqing Wu et al., "Automatic speech emotion recognition using modulation spectral features," *Speech communication*, vol. 53, no. 5, pp. 768–785, 2011.

[8] Chul Min Lee et al., "Emotion recognition based on phoneme classes," in *Proc. Interspeech*, 2004, pp. 889–892.

[9] Bogdan Vlasenko et al., "Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications," *Computer Speech & Language*, vol. 28, no. 2, pp. 483–500, 2014.

[10] Jiahong Yuan et al., "The role of phonetic units in speech emotion recognition," *arXiv preprint arXiv:2108.01132*, 2021.

[11] Rishi Bommasani et al., "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[12] Shu wen Yang et al., "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

[13] Bagus Tris Atmaja et al., "Evaluating self-supervised speech representations for speech emotion recognition," *IEEE Access*, vol. 10, pp. 124396–124407, 2022.

[14] Aaron Keesing et al., "Acoustic features and neural representations for categorical emotion recognition from speech.," in *Proc. Interspeech*, 2021, pp. 3415–3419.

[15] Sudarsana Reddy Kadiri et al., "Analysis of excitation source features of speech for emotion recognition," in *Proc. Interspeech*, 2015.

[16] J. Wagner et al., "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis amp; Machine Intelligence*, vol. 45, no. 09, pp. 10745–10759, sep 2023.

[17] Leonardo Pepino et al., "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech*, 2021, pp. 3400–3404.

[18] Zhi Zhu et al., "Deep investigation of intermediate representations in self-supervised learning models for speech emotion recognition," in *Proc. ICASSPW*, 2023.

[19] Wen Wu et al., "Estimating the Uncertainty in Emotion Attributes using Deep Evidential Regression," in *Proc. ACL*, 2023, pp. 15681–15695.

[20] Michael Grimm et al., "The Vera am Mittag German audio-visual emotional speech database," in *Proc. ICME*. IEEE, 2008, pp. 865–868.

[21] Michael Grimm et al., "Primitives-based evaluation and estimation of emotions in speech," *Speech communication*, vol. 49, no. 10-11, pp. 787–800, 2007.

[22] A. Origlia and otehrs, "Continuous emotion recognition with phonetic syllables," *Speech Communication*, vol. 57, pp. 155–169, 2014.

[23] Sadari Jayawardena et al., "How Ordinal Are Your Data?," in *Proc. Interspeech*, 2020, pp. 1853–1857.

[24] Sundararajan Srinivasan et al., "Representation learning through cross-modal conditional teacher-student training for speech emotion recognition," in *Proc. ICASSP*. IEEE, 2022, pp. 6442–6446.

[25] Bagus Tris Atmaja et al., "Improving valence prediction in dimensional speech emotion recognition using linguistic information," in *Proc. O-COCOSDA*, 2020.

[26] Srinivas Parthasarathy and Carlos Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2697–2709, 2020.

[27] Noé Tits and otehrs, "Asr-based features for emotion recognition: A transfer learning approach," *arXiv preprint arXiv:1805.09197*, 2018.

[28] Wei-Cheng Lin and Carlos Busso, "Sequential Modeling by Leveraging Non-Uniform Distribution of Speech Emotion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1087–1099, 2023.

[29] Wei-Ning Hsu et al., "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training," in *Proc. Interspeech*, 2021, pp. 721–725.

[30] Wei-Ning Hsu et al., "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.

[31] Sanyuan Chen et al., "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2021.

[32] Johannes Wagner et al., "Model for Dimensional Speech Emotion Recognition based on Wav2vec 2.0," https://doi.org/10.5281/zenodo.6221127.

[33] Carlos Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[34] Lukas Christ et al., "The MuSe 2023 Multimodal Sentiment Analysis Challenge: Mimicked Emotions, Cross-Cultural Humour, and Personalisation," in *Proc. Multimodal Sentiment Analysis Challenge (MuSe)*, 2023.

[35] Bogdan Vlasenko et al., "Fusion of acoustic and linguistic information using supervised autoencoder for improved emotion recognition," in *Proc. Multimodal Sentiment Analysis Challenge (MuSe)*, pp. 51–59. 2021.