TOWARDS LEARNING EMOTION INFORMATION FROM SHORT SEGMENTS OF SPEECH

Tilak Purohit[1,2] *Sarthak Yadav*[*3]
Bogdan Vlasenko[1] *S. Pavankumar Dubagunta*[*4] *Mathew Magimai.-Doss*[1]

[1] Idiap Research Institute, Martigny, Switzerland
[2] École polytechnique fédérale de Lausanne (EPFL), Switzerland
[3] Aalborg University, Denmark
[4] Uniphore Software Systems, India

ABSTRACT

Conventionally, speech emotion recognition has been approached by utterance or turn-level modelling of input signals, either through extracting hand-crafted low-level descriptors, bag-of-audio-words features or by feeding long-duration signals directly to deep neural networks (DNNs). While this approach has been successful, there is a growing interest in modelling speech emotion information at the short segment level, at around 250ms-500ms (e.g. the 2021-22 MuSe Challenges). This paper investigates both hand-crafted feature-based and end-to-end raw waveform DNN approaches for modelling speech emotion information in such short segments. Through experimental studies on IEMOCAP corpus, we demonstrate that the end-to-end raw waveform modelling approach is more effective than using hand-crafted features for short-segment level modelling. Furthermore, through relevance signal-based analysis of the trained neural networks, we observe that the top performing end-to-end approach tends to emphasize cepstral information instead of spectral information (such as flux and harmonicity).

Index Terms— Speech Emotion Recognition, Convolution Neural Network, End-to-End modelling

1. INTRODUCTION

In the past two decades, speech emotion recognition (SER) [1] has garnered significant attention. An important and critical aspect of SER is feature extraction. Traditional SER research was mainly devoted to the search for 'best' speech features that could provide reliable turn-level emotion classification [2, 3]. Using a brute force approach to determine the most indicative acoustic features, a large set of acoustic hand-crafted feature representations have been proposed [4–7], with feature descriptors that span intonation, intensity, cepstral-coefficients, harmonicity and perturbation-related characteristic information. However, the majority of these feature sets only model the suprasegmental nature of emotional cues. In order to show the robustness of selected frame-level and suprasegmental turn-level features, some experimental studies on speech corpora with acted emotions were reported with descriptions of the corresponding acoustic feature set like GeMAPS [6] and ComPaRE [7].

Most recent studies in the literature focus on utterance-level modelling of SER, by either extracting statistical or spectral features on the entire utterance which are then fed to standard (eg. SVMs) and deep learning-based classifiers [8–13], or by directly modelling the raw-waveform signal with deep learning approaches, such as LSTMs and TDNNs [12, 14, 15]. However, using global hand-crafted features results in a loss of fine-grained temporal information present in speech signals and it has been observed that global hand-crafted features fail to classify emotions that have similar arousal states, such as *anger* and *joy* [16]. In the case of deep learning-based techniques built on spectral features, turn-level emotion classification is typically implemented by adding a statistics pooling layer and mapping frame-level predictions obtained on standard spectral speech representations [12]. On one hand suprasegmental acoustic features representations provide discriminating information useful for detecting long-term emotional cues, on the other hand they fail to provide a deep analysis of emotion dynamics and opportunities to model paralinguistic information on the smallest possible sub-word units [17]. Furthermore, recent advances in paralinguistic speech research have moved towards modelling emotion information at the short-segment level. For instance, in the MuSE challenge [18–20], arousal and valence levels are annotated by raters at a 2 Hz sampling rate (i.e., every 500 ms) and modeled for the task of emotional stress prediction. Thus, there is a need to develop methods that can effectively model emotion information at such short segment levels.

This paper is a step in the direction of modelling sub-word segment-level speech (speech segment of about 250 ms) for speech emotion recognition. The primary research question is: how to effectively capture emotion related information at the sub-word segment level? We address this by investigating hand-crafted feature-based approach and end-to-end raw waveform modelling-based approach and demonstrate the effectiveness of the latter. A subsequent research question that arises is: what kind of emotion related information does the end-to-end raw waveform modelling based approach capture from short speech segments? To that end, besides analyzing the cumulative frequency responses of the first convolution layer and the classification performance of the evaluated methods, we also carry out speech emotion recognition studies on "relevance" signals obtained from the trained convolution neural networks and compare them to baseline speech modelling methods to gain insights into the learning dynamics of the proposed methods.

Section 2 investigates the above mentioned two approaches for short-segment modelling based SER. Section 3 presents the analysis of raw waveform CNNs, and finally, Section 4 concludes the paper.

2. SHORT SEGMENT MODELLING BASED SER

This section investigates the modelling of sub-word level short segments, viz., about 250 ms of speech for SER. As illustrated in Figure 1, we study two approaches: (1) computing the frame-level hand-

[*] work conducted while at Idiap Research Institute

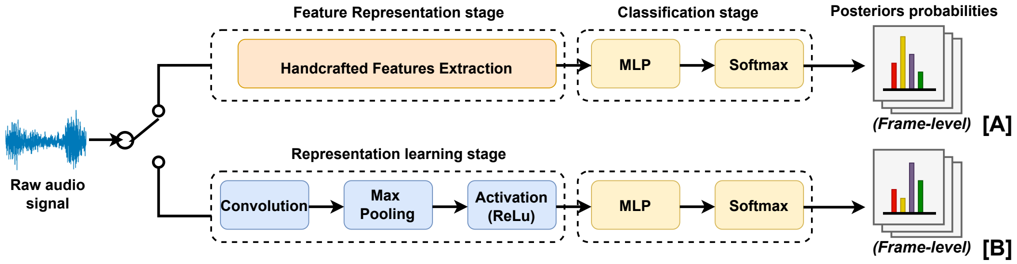

Fig. 1. Illustration of the proposed approach for modelling short segments of speech. [A] showing the approach of using handcrafted features with a short segment context. [B] showing the approach of directly modelling a short segment of raw-audio signal.

crafted features every 10 ms from the raw audio signal and feeding them as input with temporal context (12 preceding and following frames) to a multilayer perceptron (MLP) to classify emotion at frame level (Figure 1.A). (2) feeding raw audio signal of 250 ms every frame to a CNN convolution layer to classify emotion at frame level (Figure 1.B). For both, approaches the output frame level probabilities are aggregated at the utterance level to make the final decision. We study these two approaches in comparison with conventional utterance/turn-level speech segment modelling.

2.1. Database and protocols

We used the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [21], a widely used benchmark corpus in speech emotion research. IEMOCAP corpus has 12 hours of data collected from 10 subjects (5-male and 5-female) over 5 dyadic sessions. To be consistent with the previous studies [8, 22–24], we resorted to the samples from four basic emotion categories- *angry*, *happy*, *neutral* and *sad* with a total of 5531 utterances (with 1103, 1636, 1708 and 1084 utterances each, respectively) by merging the samples from the class *excited* with *happy*. Similar to the previous studies on this corpus, we conducted speaker-independent experiments following the leave-one-session-out methodology for training. For testing the k^{th} session, the model was trained on the remaining four sessions. Following the literature, the performance is measured in terms of unweighted average recall (UAR).

2.2. Conventional/ utterance-level based systems

For modelling the acoustic information for emotion classification, we decided to utilize knowledge-based feature representations provided with OPENSMILE toolkit [25] and state–of–the–art acoustic embeddings WAV2VEC2.0 [26]. The COMPARE [7] handcrafted frame-level and turn–level feature representations were used in our experimental study. Two configurations of COMPARE features were used in our experiments: COMPARE$_{LLD}$ - $65 + 65 = 130$ low–level descriptor ($LLDs$) for frame-level representation and COMPARE$_{LLD \times F}$ - 6373 static turn-level features resulting from the computation of functionals (statistics) over LLD contours. We also conducted experiments using EGEMAPS [6] 23 dimensional frame-level representations. In order to map frame-level representations into fixed–length turn–level acoustic feature vectors, we use the Bag-of-Audio-Words (BOAW) approach implemented in the OPENXBOW toolkit [27]. These features have successfully been applied for various speech applications such as acoustic event detection and speech–based emotion recognition [28, 29]. Audio chunks are represented as histograms of acoustic $LLDs$, after quantization based on a codebook. In our experimental study three configurations of BOAW were used: $500 + 500 = 1000$ codebooks for BOAW(COMPARE$_{LLD}$) (500 codebook vectors each for 65 $LLDs$ and their delta coefficients) and 1000 codebook vectors for BOAW(EGEMAPS) representation. We also built the BOAW(WAV2VEC2) system based on 768 dimensional wav2vec2.0 [26] features obtained from raw speech, using 500 codebook vectors. Further, the speech emotion classification task was carried out using these fixed–length turn–level acoustic feature vector representations by training support vector machine (SVM) and random forest (RF) classifiers.

2.3. Short-segment based systems

Handcrafted feature-based modelling: For this study we resorted to COMPARE$_{LLD}$ and EGEMAPS frame-level feature representations, consisting of feature dimension 130 and 23 respectively. Using the frame-level features we created short-segments of handcrafted features with a context of 250ms. Each frame in COMPARE$_{LLD}$ and EGEMAPS is based on a 10ms analysis window and a context of 12 preceding and succeeding frames for a total of 25 frames. These handcrafted temporal-context based features are used as input to the MLP. The number of layers and hidden nodes was decided based on the cross-validation set.

Raw audio signal modelling: We employed a raw waveform modelling approach previously proposed and studied for speech recognition [30], speaker verification [31], gender recognition [32] and depression detection [33], where raw waveform is passed through convolutional layers and then fed to an MLP for classification. We used the same architecture (4 convolutional layers followed by a single hidden layer MLP) and hyper-parameters as used for the depression detection study [33]. Depending upon the kernel width of the first convolutional layer, we distinguish two CNNs: (a) a kernel width of about 1.8 ms (< 1 pitch period) denoted as Raw SubSeg and (b) a kernel width of about 18 ms (1-5 pitch periods) denoted as Raw Seg.

To train these neural network based systems, we take an 80:20 split of each fold's training data to get train and cross validation sets. The networks were trained using cross-entropy loss with stochastic gradient descent. The learning rate was halved, in the range 10^{-1} to 10^{-6}, between successive epochs whenever the validation-loss stopped reducing. We used Keras deep learning library with tensorflow backend.

2.4. Results

Table 1 presents the performance of the different systems in terms of unweighted average recall (UAR). Table 2 presents different neural network results reported on the same protocol. In Table 1, it can be observed that the end-to-end approach yields a UAR of 57.48

Table 1. Performance of different systems on raw-audio signal, measured in terms of UAR.

Systems	Classifier	UAR
Utterance level modelling		
ComParE$_{LLD\times F}$	SVM	56.57
ComParE$_{LLD\times F}$	RF	58.23
BoAW(ComParE$_{LLD}$)	SVM	56.63
BoAW(ComParE$_{LLD}$)	RF	57.71
BoAW(eGeMAPS)	SVM	55.40
BoAW(eGeMAPS)	RF	55.90
BoAW(Wav2vec2)	SVM	53.7
BoAW(Wav2vec2)	RF	56.0
Short-segment level modelling		
ComParE$_{LLD}$	MLP	45.88
eGeMAPS	MLP	44.36
Raw SubSeg	CNN-MLP	57.48
Raw Seg	CNN-MLP	52.32

and 52.32 for Raw-SubSeg and Seg systems respectively and outperforms the hand-crafted feature-based approach which yields a UAR of 45.88 and 44.36 when modelling short speech segment. It can also be noted that the proposed short-segment level modelling end-to-end approach yields performance competitive to conventional utterance-level modeling of speech segments. It is worth mentioning that the utterance level results for ComPaRE features and BoAW word representations are comparable to those reported in literature [34]. It is interesting to note that the proposed CNN-based raw waveform modelling approach outperforms similar recent approaches that model long speech segments (from Table 2). The hand-crafted feature based systems in Table 1 and Table 2 together suggest that hand-crafted feature based approach need long segments for optimal performance. Together, these results demonstrate that the proposed end-to-end approach is able to effectively model emotion discriminating information in 250 ms of speech.

Table 2. Performance of previously reported systems measured in terms of UAR and Weighted Accuracy (WA); Utterance level (UL)

Method (Feature) – Duration	Metric	%
Att. CNN (logMel) – 7.5s [9]	WA	56.1
DBN-ivector (MFCC) – UL [13]	WA	57.2
CNN+LSTM (raw aud.) – 6s [14]	UAR	52.8
TDNN (MFCC) – 4s [15]	UAR	58.6

3. ANALYSIS OF SHORT SEGMENT MODELLING RAW CNNS

Following the results from the previous section we observe that CNNs yield competitive systems for modelling raw audio signals. Therefore, it becomes interesting to analyse what information is being learned from the 250 ms audio signal. To investigate this we conduct a three-tiered analysis: (1) We inspect the cumulative frequency response of the first CNN layer; (2) We generate relevance signals using gradient-based methods and through relevance signals, we probe what features are getting learned; and (3) Finally, we conduct an analysis at the output level, by evaluating the classification performance of selected systems.

3.1. First CNN layer frequency response analysis

Figure 2 shows the cumulative frequency response of the first convolution layer [30, 31] of the Raw-CNN for both SubSeg and Seg architectures for all the five folds M-1 to M-5. It's worth noting that both the SubSeg and Seg CNNs emphasize similar frequency regions irrespective of the fold on which they're trained on. Filters from the Raw-CNN Subseg (Fig. 2, left) emphasize the 1000 - 4000 Hz frequency region, similar to previous work where CNNs are trained to classify phones [30]. This suggests that the Raw-CNN SubSeg model is focusing on emotion related information carried at the sub-word unit level which is in line with the previous findings from [17]. For Raw-CNN Seg (Fig. 2, right), the emphasis shifts towards lower frequency regions. This observation is consistent with speaker verification and depression detection studies [31, 33], where the emphasis is on modelling voice source-related information.

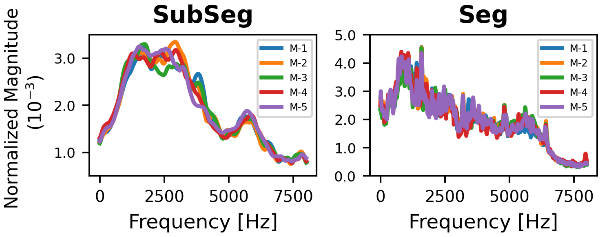

Fig. 2. Cumulative frequency response of the first convolutional layer for the proposed Raw-CNN models SubSeg (left) and Seg (right). $M - x$ indicates fold x.

The analysis together with the results obtained (in Table 1) shows that SubSeg kernel width helps in better modelling emotion class discrimination. This may be attributed to its ability to model both source and system information well when compared to Seg [35].

3.2. Relevance signal analysis

Several recent works have proposed gradient-based methods for holistic interpretation of deep feature representations, including guided-backpropagation [36] inspired gradient-based relevance signals [35, 37]. By taking the gradient of the output class with respect to the input signal, relevance signals allow us to measure the impact of perturbations in the input on the output, highlighting crucial discriminative cues in the input. We use relevance signals to gain insights into the information modeled by the proposed methods and show empirical evidence to support our hypothesis.

Table 3. Performance of different systems on relevance signal, measured in terms of UAR.

Systems	Input Signal	Classifier	UAR
Utterance level modelling			
ComParE$_{LLD\times F}$	SubSeg-Rel	SVM	57.15
ComParE$_{LLD\times F}$	SubSeg-Rel	RF	54.06
ComParE$_{LLD\times F}$	Seg-Rel	SVM	50.57
ComParE$_{LLD\times F}$	Seg-Rel	RF	54.62
Short-segment level modelling			
Raw-CNN SubSeg	SubSeg-Rel	Softmax	56.37
Raw-CNN Seg	Seg-Rel	Softmax	49.96

To ascertain that relevance signals indeed represent crucial discriminative cues in the input necessary for emotion recognition, we train both the proposed raw CNN models on short-segment relevance signals to classify emotion recognition. Using the respective trained CNN, we compute relevance signals for every input in the training data with respect to the ground truth, denoted by SubSeg-Rel and Seg-Rel corresponding to both the proposed SubSeg and Seg CNNs. These relevance signals (SubSeg-Rel and Seg-Rel) were used for training the models. At test time, instead of computing relevance signal only for the ground truth, we generate relevance signal for all four classes and average the predictions of the model on all these four relevance signals. This process is repeated for each fold and UAR is reported, as seen in the bottom two rows of Table 3. We can see that each of the models achieves performance quite close to the original SubSeg and Seg models trained on raw audio signals (from Table 1). We repeat this procedure for utterance-level modelling by training SVM and RF classifiers on ComParE$_{LLD}$ features computed from utterance-level relevance signals. From Table 3, we can see an improvement in the performance of the SVM classifier when trained using relevance signals obtained from the SubSeg model over the original raw waveform signal (Table 1). Together, these results indicate that relevance signals indeed represent information in the input crucial for emotion recognition.

To get insights into the information modeled by the proposed CNNs, we rank the top-10 features based on normalized feature importance assigned by RF classifiers trained on ComParE$_{LLD}$ feature descriptors on the original raw-waveform signal and the relevance signals obtained from the SubSeg and Seg CNNs, as shown in Table 4. For the sake of clarity, full feature names were omitted from the table, and the F-Index column indicates the i-th feature from the 0-indexed feature list from the ComParE header extracted from openSmile [25] toolkit. The “Group” column highlights the broader feature group of the low-level descriptor, the feature grouping has been adopted from [38]. In general, the raw-waveform model primarily focuses on spectral low-level feature descriptors (9/10), primarily spectral flux and harmonicity. The SubSeg-Rel based model primarily focuses on cepstral feature descriptors (8/10), more so than the Seg-Rel based model (4/10). It's worth pointing out that the Seg-Rel model puts a lot of emphasis on spectral features (6/10), similar to the raw-waveform model, focusing more on spectral slope descriptors instead of harmonicity. Overall, this section further highlights the different modelling characteristics of the two proposed CNNs.

Raw-waveform		SubSeg-Rel		Seg-Rel	
F-Index	Group	F-Index	Group	F-Index	Group
2957	Spectral Flux	5168	Cepstral	5166	Cepstral
1513	Spectral Harmonicity	5166	Cepstral	1523	Cepstral
4138	Spectral (Auditory)	1637	Cepstral	1522	Cepstral
3218	Spectral Harmonicity	1522	Cepstral	2957	Spectral Flux
5496	Spectral (Auditory)	1523	Cepstral	1431	Spectral Slope
6039	Spectral Flux	1587	Cepstral	5097	Spectral Slope
5490	Spectral (Auditory)	5173	Cepstral	5489	Spectral (Auditory)
6035	Spectral Flux	3311	Cepstral	132	Spectral (Auditory)
74	Prosodic	1244	Spectral Flux	5173	Cepstral
6030	Spectral Flux	3712	Voice quality	4144	Spectral (Auditory)

Table 4. Ranking feature importances from utterance-level RF classifiers trained on ComParE$_{LLD}$ features obtained from different input signals. Feature groups as per [38]

3.3. Classification analysis

Figure 3 presents the confusion matrices of the proposed SubSeg raw CNN trained on raw-audio (Fig 3.c) and relevance signals (Fig 3.d). For comparison, confusion matrices of corresponding utterance-level ComParE$_{LLD}$ systems are also shown (Fig 3.a and 3.b, respectively). We can observe that modelling emotions at sub-word unit level with the SubSeg model provides a better classification performance for happy and sad emotional states as compared to ComParE$_{LLD}$ based systems. Also, for the SubSeg model the emotion state that achieves the highest UAR varies across input signals (sad and happy for raw audio and SubSeg-Rel, respectively). In contrast to raw-audio based methods, ComParE$_{LLD}$ based systems class-wise recall rates are better for the neutral and angry state. It's also worth noting that when the SubSeg-Rel input is used, ComParE$_{LLD}$ system models all emotional states equally.

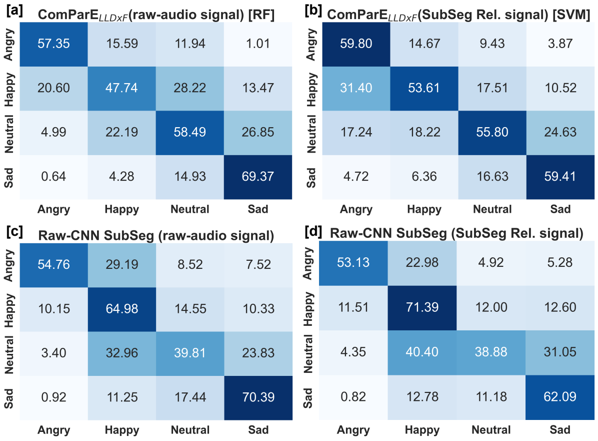

Fig. 3. Confusion matrices of four selected systems : [a],[b] systems build on knowledge-based acoustic features and [c],[d] build on proposed raw-CNN system

4. CONCLUSION

In this paper, we investigated two approaches, namely (a) handcrafted feature-based and (b) end-to-end raw waveform DNNs, for modelling speech emotion information in short segment speech of 250 ms duration. Experimental studies on the IEMOCAP corpus demonstrated that the end-to-end SubSeg modelling approach is able to achieve performance on par with conventional utterance/turn-level modelling of longer speech segments (4+ seconds of speech), despite being trained on such short-segment level input whereas, modelling similar short-segments of handcrafted features does not. SER studies carried out on relevance signals computed from trained CNNs revealed that SubSeg modelling tends to prioritize cepstral information while Seg modelling emphasizes both cepstral and spectral information. Finally, analysis of classifier outputs showed that the end-to-end approach is able to discriminate high and low arousal states well. Our future work will focus on extending the investigation to MuSe emotional stress sub-challenge [18, 19], where arousal and valence are modeled at a continuous level.

5. 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. $40B2 - 0\_194794$). This research was partially supported by the TPU Research Cloud (TRC) program, a Google Research initiative.

6. REFERENCES

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

[2] B.W. Schuller 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] Mo. El Ayadi et al., "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition*, 2011.

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

[6] F. Eyben et al., "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, 2016.

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

[8] S. Ghosh et al., "Representation learning for speech emotion recognition.," in *Proc. of Interspeech*, 2016.

[9] M. Neumann and 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.

[10] E. Kim and J. W. Shin, "DNN-based emotion recognition based on bottleneck acoustic features and lexical features," in *Proc. of ICASSP*, 2019.

[11] Z. Peng et al., "Efficient speech emotion recognition using multi-scale CNN and attention," in *Proc. of ICASSP*, 2021.

[12] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, 2019.

[13] R. Xia and Y. Liu, "DBN-ivector Framework for Acoustic Emotion Recognition," in *Proc. Interspeech*, 2016.

[14] J.L. Li et al., "A waveform-feature dual branch acoustic embedding network for emotion recognition," *Frontiers in Computer Science*, 2020.

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

[16] T.L. Nwe, S.W.Foo, and L.C De Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, 2003.

[17] B. Vlasenko and A. Wendemuth, "Determining the smallest emotional unit for level of arousal classification," in *Proc. of ACII*, 2013.

[18] L. Stappen et al., "MuSe 2021 Challenge: Multimodal Emotion, Sentiment, Physiological-Emotion, and Stress Detection," in *Proc. of ACM Multimedia*, 2021.

[19] S. Amiriparian et al., "MuSe 2022 Challenge: Multimodal Humour, Emotional Reactions, and Stress," in *Proc. of ACM Multimedia*, 2022.

[20] S. Yadav et al., "Comparing biosignal and acoustic feature representation for continuous emotion recognition," in *Proc. of International Multimodal Sentiment Analysis Workshop and Challenge*, 2022.

[21] C. Busso and other, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, 2008.

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

[23] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2D continuous space," *IEEE Transactions on affective computing*, 2015.

[24] V. Rozgić et al., "Ensemble of SVM trees for multimodal emotion recognition," in *Proceedings Asia Pacific Signal and Information processing Association Annual Summit and Conference*, 2012.

[25] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. of ACM Multimedia*, 2010.

[26] A Baevski et al., "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Information Processing Systems*, 2020.

[27] M. Schmitt and Bjőrn Schuller, "openXBOW - Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit," *Journal of Machine Learning Research*, 2017.

[28] H. Lim, M. J. Kim, and H. Kim, "Robust sound event classification using LBP-HOG based bag-of-audio-words feature representation," in *Proc. of Interspeech*, 2015.

[29] M. Schmitt, F. Ringeval, and B. W. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech.," in *Proc. of Interspeech*, 2016.

[30] D. Palaz, M. Magimai.-Doss, and R. Collobert, "End-to-End Acoustic Modeling using Convolutional Neural Networks for HMM-based Automatic Speech Recognition," *Speech Communication*, 2019.

[31] H. Muckenhirn, M. Magimai.-Doss, and S. Marcel, "Towards directly modeling raw speech signal for speaker verification using CNNs," in *Proc. of ICASSP*, 2018.

[32] S. H. Kabil, H. Muckenhirn, and M. Magimai-Doss, "On learning to identify genders from raw speech signal using CNNs," in *Proc. of Interspeech*, 2018.

[33] S. P. Dubagunta, B. Vlasenko, and M. Magimai.-Doss, "Learning voice source related information for depression detection," in *Proc. of ICASSP*, 2019.

[34] S. Amiriparian 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] H. Muckenhirn et al., "Understanding and visualizing raw waveform-based CNNs," in *Proc. of Interspeech*, 2019.

[36] J. T. Springenberg et al., "Striving for simplicity: The all convolutional net," in *ICLR (Workshop)*, 2015.

[37] H. Muckenhirn et al., "Gradient-based spectral visualization of CNNs using raw waveforms," Tech. Rep., Idiap, 2018.

[38] B.W. Schuller et al., "The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load, multitasking," in *Proc. of Interspeech*, 2014.