Predicting Arousal and Valence from Waveforms and Spectrograms using Deep Neural Networks

Zixiaofan Yang[1]*, Julia Hirschberg*[1]

[1]Columbia University

zy2231@columbia.edu, julia@cs.columbia.edu

Abstract

Automatic recognition of spontaneous emotion in conversational speech is an important yet challenging problem. In this paper, we propose a deep neural network model to track continuous emotion changes in the arousal-valence two-dimensional space by combining inputs from raw waveform signals and spectrograms, both of which have been shown to be useful in the emotion recognition task. The neural network architecture contains a set of convolutional neural network (CNN) layers and bidirectional long short-term memory (BLSTM) layers to account for both temporal and spectral variation and model contextual content. Experimental results of predicting valence and arousal on the SEMAINE database and the RECOLA database show that the proposed model significantly outperforms model using hand-engineered features, by exploiting waveforms and spectrograms as input. We also compare the effects of waveforms vs. spectrograms and find that waveforms are better at capturing arousal, while spectrograms are better at capturing valence. Moreover, combining information from both inputs provides further improvement to the performance.

Index Terms: speech emotion recognition, computational paralinguistics, deep learning

1. Introduction

In recent years, increasing attention has been given to the study of the emotional content in speech signals, and many systems have been proposed for automatic emotion recognition in speech. For most systems, the goal is to produce a categorical label among a set of 'basic emotions' such as disgust, sadness, happiness, fear, anger and surprise [1]. This view of emotion originates in expressions in human language describing emotional experiences in terms of words [2]. However, speech signals contain more subtle changes in emotion, especially for conversational speech and spontaneous emotion in which both speakers' affective states change continuously over time. In this case, a categorical approach may fail to capture changes. Also, some emotions are easier to distinguish, while others share similar characteristics [3]. The similarity/disparity issue among emotion categories also represents a potential problem in automatic emotion classification. However, another fundamental approach to emotion detection is to map emotion into a continuous multi-dimensional space. The underlying assumption in this approach is that a common physiological system is responsible for all emotional states. When measuring emotion using this dimensional approach, the emotion recognition task can be treated as a regression problem. In each of the dimensions, we can use a series of float numbers to represent the target's emotion. One of the most prominent models taking this viewpoint is Russell's circumplex model of emotion [4]. In the circumplex model, a person's emotion is described as a point in the arousal-valence two-dimensional space. Predicting continuously changing arousal and valence is inherently a more difficult task than classifying discrete emotions for each utterance, due to its high granularity in both the emotion domain and the time domain. However, this approach to emotion detection can better represent natural speech in real situations. Our work follows the circumplex model and our goal is to produce numerical predictions for both arousal and valence from speech.

In traditional methods of emotional speech recognition, features are hand-engineered, selected using prior knowledge of the auditory signal processing area, such as pitch, intensity, speaking rate and mel-frequency cepstral coefficients (MFCC) [5]. However, recent advances in computing resources and neural network architectures have enabled end-to-end speech processing, in which inputs are drawn directly from minimally processed speech data such as waveforms and spectrograms [6, 7, 8]. In recognizing emotional speech, mel-scale filterbank spectrograms are widely used as input features to neural network models because of their close relationship with human perception of speech signals [9]. Also, recent research has shown that neural networks can automatically learn some emotion-related feature representations such as energy and fundamental frequency from raw waveform signals [10]. However, there is currently no work exploring whether waveforms and spectrograms also contain complementary information on emotional speech. In this work, we combine inputs from raw waveform signals and mel-scale log filter-bank features to examine their joint effects. The neural network architecture that we use contains a set of convolutional neural network (CNN) layers and bidirectional long short-term memory (BLSTM) layers to account for both temporal and spectral variation and model contextual content.

The paper is organized as follows. Section 2 summarizes prior work on emotion recognition from speech. In Section 4, we introduce the input features and the topology of the neural network architecture. The corpora used are described in Section 3, and our experimental results are presented in Section 5. Finally, in Section 6 we conclude and present future work.

2. Related Work

There has been considerable research on improving neural network structures for emotion recognition in speech. For most such research, the goal is to predict a label among a fixed set of discrete emotions. Han et al. [11] proposed a deep neural network and extreme learning machine (DNN-ELM) model to recognize excitement, frustration, happiness, neutral and surprise. Mao et al. [12] used a CNN to learn affect-salient features from spectrograms. In the experiments, they used 4 corpora with four different sets of emotions, including: (1) anger, disgust, fear, happiness, sadness, surprise, and neutral; (2) anger, disgust, fear, joy, sadness, boredom and neutral; (3) anger, joy, surprise, sadness and neutral; (4) anger, joy, surprise, sadness

and disgust. Lee et al. [13] used RNN on frame-level hand-engineered features to recognize happiness, sadness, anger and neutral. Recently, Mirsamadi et al. [14] used RNNs with an attention mechanism to focus on emotionally salient regions for happiness, sadness, anger and neutral. Huang and Narayanan [15] used CNN-LSTM-DNNs with an attention mechanism to classify anger, disgust, fear, joy, sadness, and surprise. Kim et al. explored the effect of 3D CNNs [16] and skip-connections [17] on happiness, sadness, anger and neutral. Cummins et al. [18] used pre-trained image classification CNN to process spectrograms and recognize angry, emphatic, neutral, postive and rest. Finally, Bertero and Fung [19] found that their CNN filters concentrated on the average pitch range related to emotions such as as angry, happy and sad on the frequency domain and activated during the speech sections while ignoring the silence parts on time domain. In the work discussed above, a total of 8 different sets of discrete emotions are used, which makes it difficult to compare models optimized for different emotions.

There is also research on predicting continuous emotion in the arousal-valence two-dimensional space. Giannakopoulos et al. [20] conducted emotion recognition in arousal-valence space and found that this approach offers a good affective representation for speech. Towards better feature representations, Schmitt et al. [21] explored bag-of-audio-words representation of MFCCs as input to the regression model, and Zhang et al. [22] performed feature enhancement using an autoencoder with LSTM. Towards better neural network structures, Trigeorgis et al. [10] proposed an CNN-LSTM-DNN on waveform signals, and Han et al. [23] concatenated different regression models to exploit their individual advantages. However, little existing work has explored the difference in predicting valence and arousal [24].

3. Corpora

To evaluate the performance of our model, we need speech corpora with continuous annotations of arousal and valence on a high granularity. For this purpose, we chose two corpora of natural conversational speech: the SEMAINE database [25] and the RECOLA database[26].

3.1. The SEMAINE database

The SEMAINE database was collected to study emotionally colored conversations in English and has the highest annotation granularity of all publicly-available corpora. In SEMAINE recording, two speakers in each conversation are a user and an 'operator' who simulates a Sensitive Artificial Listener (SAL) agent. The goal of the operator is to engage the user in emotional conversations by asking questions and expressing attitudes, such as 'Anything else nice happened this week?' or 'It is all rubbish.' To ensure that we are looking at truly spontaneous emotions in speech, we use only the Solid-SAL session with the most natural operator interactions, and look only at the user's turns from each conversation. The user's emotion is annotated by 6-8 annotators for arousal and valence at 20ms intervals; annotation scores range from -1 to 1 with 4 decimal places. We segment the 83 conversations with 24 users into turns according to the transcripts, aligning the user turns with the averaged manual annotations. We randomly employ 70% of the conversations with 934 6s segments as the training set, and the remaining 30% with 396 6s segments as the test set.

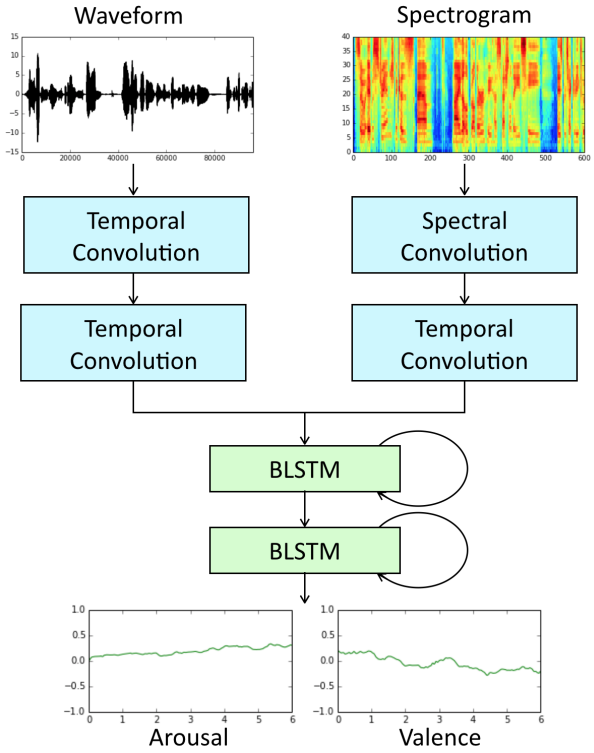

Figure 1: *The architecture of the proposed model.*

3.2. The RECOLA database

The RECOLA database is a multi-modal corpus of spontaneous collaborative and affective interaction in French. After completing a self-report questionnaire, 46 subjects watched video clips for positive/negative mood manipulation and then participated in a task in which they were asked to reach consensus on how to survive in a disaster scenario. This task was intended to trigger emotional communication between participants. Conversations were annotated for arousal and valence at 40ms intervals by 6 annotators; scores range from -1 to 1 with 2 decimal places. The version we employ contains 23 conversations, each lasting 5 minutes. Since both speakers show spontaneous emotions and turn-taking information is not provided, we use entire conversations without segmenting speaker turns. As with the SEMAINE database, we randomly use 70% with 800 6s segments for training and 30% with 350 6s segments for testing.

4. Model

We use an end-to-end deep convolutional recurrent neural network to perform emotion recognition; the architecture of this network is shown in Figure 1. The main difference from a standard CNN-LSTM-DNN architecture is that two sets of 1-D CNN layers are used separately to process two types of raw features we believe to contain complementary information about arousal and valence. The output of these CNN layers are then concatenated together and fed into the BLSTM layers to generate the prediction of both arousal and valence. The CNN layers can reduce temporal and spectral variation and exploit the information contained in the two inputs, while the BLSTM layers can take contextual content into account and generate predictions with high temporal granularity.

4.1. Input Data

4.1.1. Raw waveform signals

With the use of deep neural network structures, raw waveform signals have been shown to be useful in numerous speech recognition tasks, providing information such as loudness, energy and pitch. For preprocessing, we normalize waveform signals on the conversation level with zero mean and unit variance to reduce the inter-speaker difference. Then we re-sample the speech to 16kHz sampling rate, and segment the conversation into 6s segments with 96,000 samples as the waveform input. An example of the raw waveform signals is shown at the upper left corner of Figure 1.

4.1.2. Spectrogram features

Previous studies have found that the waveforms and the spectrograms provide complementary information in learning acoustic models [8]. These findings have inspired us to include spectrograms as another input to our neural network. We use the output of a 40-dimensional mel-scale log filter bank as the spectrogram features. Similar with our preprocessing of waveforms, we first perform normalization and segmentation. The spectrogram features and the first and second temporal derivatives are then computed over windows of 25ms length and 10ms stride, resulting in three 40*600 matrices for each 6s segment. An example of these spectrogram features is shown at the upper right corner of Figure 1. The horizontal axis represents time in frames, and the vertical axis represents filter banks with different frequency ranges. For display purpose, the temporal derivatives are not shown in this figure.

4.2. Neural Network Architecture

4.2.1. CNN layers

For the waveform input, the CNN layers are used to extract information in different temporal scales. The first layer has 40 channels with a kernel size of 80, followed by a max pooling layer with size of 2. The second layer has a kernel size of 800, followed by a cross-channel max pooling layer with a size of 20. The convolution filter in the first layer has a receptive field of 5ms, while the filter in the second layer has a receptive filed of 100ms. In this way, the two CNN layers can jointly learn frame-level features as well as long-term patterns.

For the spectrogram input, the CNN layers are used to reduce temporal and spectral variation while preserving locality. The first layer is a spectral convolution layer. It has 80 channels with a kernel size of 10, followed by a spectral max pooling layer with size of 2. The second layer is a temporal convolution layer. It has a kernel size of 10, followed by a cross-channel max pooling layer with size of 10. The temporal convolution filter for the spectrogram input has a receptive field of 115ms which is roughly the same as the waveform input in order to extract long-term patterns on a similar scale.

4.2.2. BLSTM layers

Both of the CNN layers produce 96000-dimensional output vectors from the 6s inputs of waveforms and spectrograms. The CNN output vectors are segmented into millisecond-level pieces depending on the granularity of the annotations and concatenated together (e.g. two 320-dimensional pieces for 20ms annotations) to feed into the BLSTM layers. We use two BLSTM layers with 256 cells each to further reduce temporal variation and model contextual information. Finally, a fully connected layer follows each output of BLSTM to generate the numerical predictions of arousal and valence.

Corpus	Model	Results (CCC)	
		Arousal	Valence
SEMAINE	Baseline	0.376	0.177
	W Only	*0.675*	0.435
	S Only	0.656	*0.494*
	W + S	**0.680**	**0.506**
RECOLA	Baseline	0.317	0.162
	W Only	*0.674*	0.361
	S Only	0.651	*0.408*
	W + S	**0.692**	**0.423**

Table 1: *The concordance correlation coefficient (CCC) of the baseline model and three proposed models on the SEMAINE database and the RECOLA database.*

5. Experiments

5.1. Overview

For our experiments on the two datasets, we first implement a baseline model with hand-engineered features and BLSTM layers. We use the openSMILE toolkit [27] to extract the ComParE feature set [28] with 6373 features, which is the official baseline set for the INTERSPEECH ComParE challenges from 2013 to 2017. The hand-engineered features are extracted on a 1s window with the same temporal stride as the annotations. Then, to compare the difference between waveform and spectrogram inputs, we create three end-to-end models, one using only waveform input ('W Only'), one using only spectrogram input ('S Only'), and a third combining both inputs ('W+S'). To make the comparison fair, the BLSTM layers of the 'W Only' and 'S Only' models have half the number of cells as the 'W+S' model.

For all these experiments, we use the concordance correlation coefficient (CCC) [29] as the objective function to train the models. CCC measures the similarity between two sequences of numbers, a metric which is commonly used in continuous emotion recognition task. All the neural network models are trained with a RMSProp optimizer with a learning rate of $5*10^{-4}$ and a batch size of 50. All CNN layers use ReLU activation. Dropout layers with 0.5 dropout rate are added after the max-pooling layers.

5.2. Results

The experimental results on the SEMAINE database and the RECOLA database are shown in Table 1. Firstly, all our models perform significantly better than the baseline model, which indicates that the models can learn salient features for arousal and valence from either of the inputs. Moreover, in both of the corpora, the 'W Only' model outperforms the 'S Only' model in predicting arousal, while the 'S Only' model outperforms the 'W Only' model in predicting valence. This might be explained by: (1) The fact that the arousal dimension is related to the 'loudness' of the speech, and the root-mean-square amplitude for acoustic intensity can be directly extracted from the waveform. (2) The valence dimension is more complex and cannot be easily related to any particular speech characteristics. However, the spectrograms offer more interpretability with respect to articulation and pitch, and thus allow the model to learn patterns from a spectral aspect. Finally, combining both the waveform

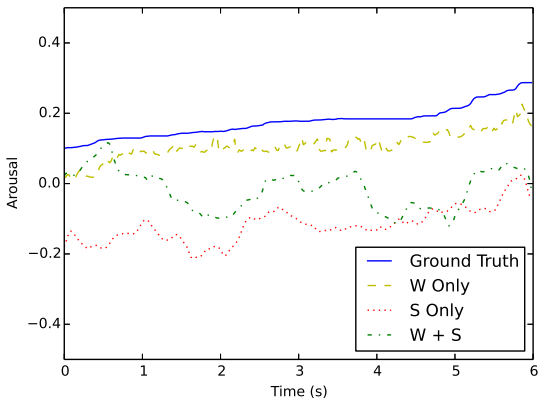

Figure 2: *Predictions of arousal on an instance.*

and the spectrogram inputs, the 'W + S' model provides further improvement in predicting both arousal and valence, which demonstrates that waveforms and spectrograms do contain complementary information of emotion. Comparing results for the two corpora, the CCC for predicting valence on SEMAINE is systematically higher than that on RECOLA. This may be because of the different strategies for inducing emotional conversations. The operator in SEMAINE tends to induce extreme values on valence, which makes the variance of valence 1.55 times larger than the variance of arousal. In RECOLA, the two speakers are communicating after the positive/negative mood induction procedure, and the variances of arousal and valence are roughly the same. Our results are comparable to state-of-the-art results on RECOLA with a CCC of 0.744 for arousal and 0.393 for valence [23], although this study used the full dataset of 46 conversations while we could only obtain 23 of them. The best results on the SEMAINE database reported Mean Correlation Coefficient scores (MCC) for arousal 0.521 and valence 0.211 [30], while our 'W + S' model obtains MCC for arousal 0.682 and valence 0.511 on the test set.

5.3. Analysis

Figure 2 and Figure 3 show the ground truth and the predictions of our three models on a segment of the SEMAINE database. The solid blue line represents ground truth, the dashed yellow line is the output of the 'W Only' model, the dotted red line is the output of the 'S Only' model and the green line with both dash and dot is the output of the combined 'W + S' model. The transcript of the speech segment is 'Ehh.... of all the characters, Prudence is the one who gets under my skin, cos she's so frigging superior.' From Figure 2, we observe that the 'W Only' model performs the best with correct polarity and trend, and the 'S Only' model predicts the wrong arousal polarity. From Figure 3, we observe that the 'S Only' model captures the descending trend while the 'W Only' model does not capture it. We also find that the sudden drop in 'S Only' output at around 4.7s matches the time of the word 'frigging', which is used here to emphasize negative valence. To examine the effect of spectrogram input towards the output crest at 4.7s, we employ a novel method called the Local Interpretable Modelagnostic Explanations (LIME) [31], which has not yet been applied to any speech recognition model. Since spectrograms share dimensional and locality similarity with images, we use the image explanation

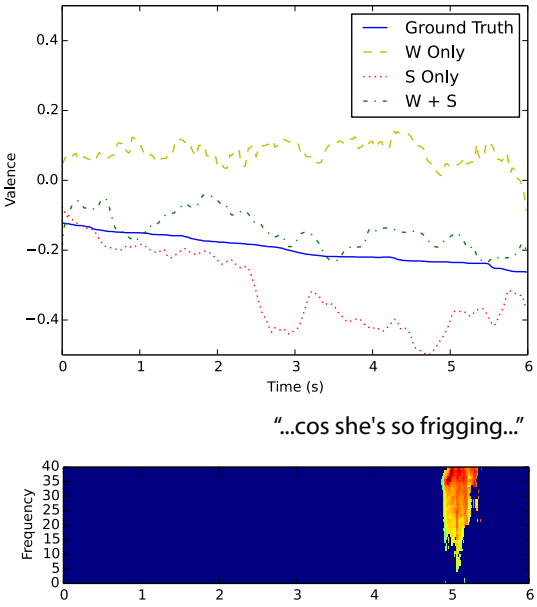

Figure 3: *The upper part is the prediction of valence on an instance. The lower part is the LIME explanation of the crest in 'S Only' output.*

module of LIME; the explanation of the output crest is shown in the lower part of 3. The most important part of the spectrogram input for the crest is highlighted with bright colors, while the other parts remain dark blue. The LIME explanation shows that the high energy of the higher frequency components from around 4.9s to 5.3s leads to the drop in valence prediction at around 4.7s. Using the LIME method, we can also generate explanations for other instances to better understand the performance of the models.

6. Conclusions and Future Work

We propose a deep convolutional recurrent network model to predict arousal and valence by combining inputs from raw waveform signals and spectrogram features. We conducted experiments on the SEMAINE and the RECOLA corpora, and our models significantly outperform hand-engineered features. By comparing the models with waveforms only and spectrograms only, we found that waveforms are better at capturing arousal, spectrograms are better at valence, and combining both provides further improvement. We also analyzed an instance using LIME to better understand the model. In future, we plan to perform deeper analysis of the inputs to further exploit their strength. We are also interested in building models that can assign different weights to the inputs according to the characteristics of the instance.

7. Acknowledgements

This work was funded by DARPA LORELEI grant HR0011-15-2-0041. The views expressed in this paper however are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S government.

8. References

[1] P. Ekman, W. V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion." *Journal of personality and social psychology*, vol. 53, no. 4, p. 712, 1987.

[2] K. R. Scherer, "What are emotions? and how can they be measured?" *Social science information*, vol. 44, no. 4, pp. 695–729, 2005.

[3] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression." *Journal of personality and social psychology*, vol. 70, no. 3, p. 614, 1996.

[4] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant." *Journal of personality and social psychology*, vol. 76, no. 5, p. 805, 1999.

[5] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[6] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, 2013, pp. 6645–6649.

[7] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4580–4584.

[8] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform cldnns," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[9] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *Readings in speech recognition*. Elsevier, 1990, pp. 65–74.

[10] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5200–5204.

[11] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[12] 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.

[13] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[14] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2227–2231.

[15] C.-W. Huang and S. S. Narayanan, "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition," in *Multimedia and Expo (ICME), 2017 IEEE International Conference on*. IEEE, 2017, pp. 583–588.

[16] J. Kim, K. P. Truong, G. Englebienne, and V. Evers, "Learning spectro-temporal features with 3d cnns for speech emotion recognition," *arXiv preprint arXiv:1708.05071*, 2017.

[17] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, "Deep temporal models using identity skip-connections for speech emotion recognition," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 1006–1013.

[18] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 478–484.

[19] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5115–5119.

[20] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, "A dimensional approach to emotion recognition of speech from movies," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 65–68.

[21] 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 *INTERSPEECH*, 2016, pp. 495–499.

[22] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller, "Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with lstm neural networks," in *17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016)*, 2016, pp. 3593–3597.

[23] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Prediction-based learning for continuous emotion recognition in speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5005–5009.

[24] J. Liscombe, J. Venditti, and J. Hirschberg, "Classifying subject ratings of emotional speech using acoustic features," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[25] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[26] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[27] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[28] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom *et al.*, "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Computational Paralinguistics Challenge (ComParE), Interspeech 2017*, 2017, pp. 3442–3446.

[29] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.

[30] T. Dang, V. Sethu, and E. Ambikairajah, "Factor analysis based speaker normalisation for continuous emotion prediction." in *INTERSPEECH*, 2016, pp. 913–917.

[31] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 1135–1144.