INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

A Multi-scale Fusion Framework for Bimodal Speech Emotion Recognition

Ming Chen[1,2], *Xudong Zhao*[2]

[1]Zhejiang University, No.38 Zheda Road, Hangzhou, China
[2]Hithink RoyalFlush Information Network Co., Ltd., Hangzhou, China

chm@zju.edu.cn, sdu_zxd@163.com

Abstract

Speech emotion recognition (SER) is a challenging task that requires to learn suitable features for achieving good performance. The development of deep learning techniques makes it possible to automatically extract features rather than construct hand-crafted features. In this paper, a multi-scale fusion framework named STSER is proposed for bimodal SER by using speech and text information. A smodel, which takes advantage of convolutional neural network (CNN), bi-directional long short-term memory (Bi-LSTM) and the attention mechanism, is proposed to learn speech representation from the log-mel spectrogram extracted from speech data. Specifically, the CNN layers are utilized to learn local correlations. Then the Bi-LSTM layer is applied to learn long-term dependencies and contextual information. Finally, the multi-head self-attention layer makes the model focus on the features that are most related to the emotions. A tmodel using a pre-trained ALBERT model is applied for learning text representation from text data. Finally, a multi-scale fusion strategy, including feature fusion and ensemble learning, is applied to improve the overall performance. Experiments conducted on the public emotion dataset IEMOCAP have shown that the proposed STSER can achieve comparable recognition accuracy with fewer feature inputs.

Index Terms: speech emotion recognition, bimodal, multi-scale fusion strategy, feature fusion, ensemble learning

1. Introduction

Recent years have witnessed significant advances in the artificial intelligence field, the human-computer interaction (HCI), which is an important part, is being performed through a variety of softwares, smart devices and so on. Understanding user intention is an essential factor in enriching user experience and it has become a hot topic in both research and industry areas. Just as people communicate with each other, speech plays a crucial role in capturing the explicit content messages and even the underlying intentions. That is because speech contains rich linguistic and para-linguistic information conveys the implicit information such as emotions [1]. Current techniques (such as [2]) have achieved high accuracy of recognizing the content from the speech utterance; however, it is not sufficient for HCI systems to fully understand the purpose and intention of the speaker. Para-linguistic information, especially emotions, can help HCI systems capture the real purpose and right feeling of speakers, thereby giving a better response. As a result, speech emotion recognition (SER) has received increasing attention.

To determine the classification of emotions, the fundamental and general approach uses discrete states to represent emotions, such as anger, happiness and so on. Another alternative method is to use a three-dimensional continuous space (i.e., arousal, valence and potency) to represent emotions. In this paper, we focus on the first discrete approach.

There are mainly two reasons why SER is challenging:

- There exist various emotional expressions by different individuals. It is challenging to correctly classify individual emotions even with models that learn better general characteristics.
- For a long speech utterance, there are many short periods of silence, and in many cases, only some specific moments are available to detect emotions. Besides, most SER datasets provide the emotion labels at the whole speech utterance level.

To address the SER task, many existing works have been performed on designing and utilizing distinguished speech features to indicate different emotions [3]. Traditionally, a large number of hand-crafted features (such as pitch, mel-scale frequency cepstral coefficient (MFCC) and so on) are firstly extracted from raw speech data, and then a compact feature representation vector is generated by applying various statistical functions, such as min, max and so on. This feature vector is assumed to be related to the emotion of the whole speech utterance. Finally, the vector is used as the input of the machine learning algorithm for classification [4].

For these traditional approaches, the quality of the hand-crafted features dramatically affects the final classification results. Therefore, some researchers have focused on how to extract more useful and suitable features from speech signals and then take advantage of these features to improve recognition accuracy. Kim et al. exploited the extended geneva minimalistic acoustic parameter set (eGeMAPS) features of the temporal information for SER [5]. However, designing and selecting suitable features often requires professional knowledge.

Deep learning has been considered as an emerging research field in machine learning and has made significant progress in various areas, including image classification [6], speech recognition [7] and so on. Hence, more and more researchers have used deep neural networks (DNNs) for addressing SER task [8, 9]. Stuhlsatz et al. [9] used a DNN model to learn discriminative representations from traditional statistical features for SER and improved the accuracy compared with Support Vector Machines (SVM); however, a large set of hand-crafted features are required as the model input for better performance. In this paper, to use as few hand-crafted features as possible, with respect to speech data, only the log-mel spectrograms are used. Meanwhile, the corresponding text data are also utilized for further increasing the performance of SER.

In this paper, we present a multi-scale fusion framework named STSER, which utilizes both the speech and text information, for bimodal SER task. For speech data, A smodel is proposed to extract features from log-mel spectrograms by using the strength of the CNN, Bi-LSTM and the attention mechanism. Especially, CNN layers are firstly used to learn local correlations, and one Bi-LSTM layer is used for learning long-term dependencies and contextual information. Besides, as mentioned

http://dx.doi.org/10.21437/Interspeech.2020-3156

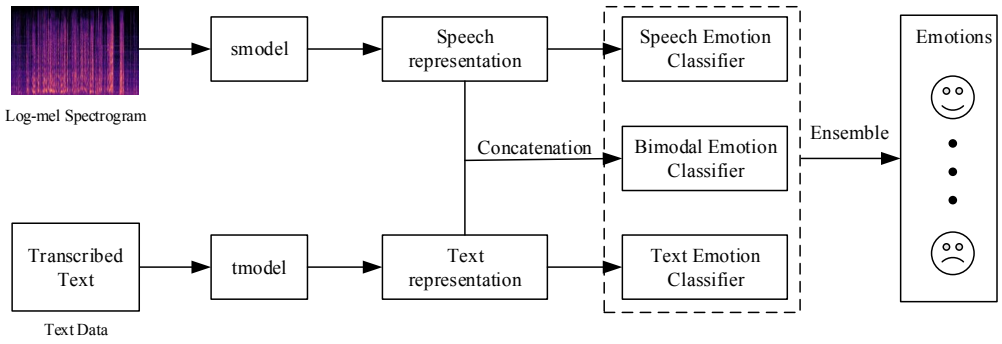

Figure 1: *Overview structure of the proposed STSER for SER*

above, only specific parts contribute to the emotion of the whole speech utterance. Therefore, to address this problem, the multi-head self-attention layer is introduced to make the smodel focus on the features related to the emotion. For the text data, a pre-trained ALBERT model [10] is applied for learning the text representation for assisting emotion recognition. Finally, a multi-scale fusion strategy, including feature fusion and ensemble learning, is utilized for improving the overall performance. The whole STSER is evaluated on the public emotion dataset - Interactive Emotional Dyadic Motion Capture database (IEMOCAP) [11]. Experimental results have shown that the proposed framework can achieve comparable recognition accuracy with fewer inputs.

The rest of the paper is structured as follows. Section 2 gives the related work. Section 3 describes the details of the proposed deep learning-based model for SER task. Section 4 presents and analyzes the experimental results of the proposed model. Section 5 concludes the paper.

2. Related Work

Speech emotion recognition has attracted a great deal of attention in the area of signal processing. With the development of deep learning techniques, especially inspired by the successful application of image processing and natural language processing (NLP), many existing researches, which are based on the DNNs, have been proposed to recognize speech emotions. For example, Wollmer et al. [12] proposed a DNN architecture consisting of a three layer LSTM for SER task. Tang et al. [13] proposed three models under the end-to-end learning framework for SER task, including CNN combined with extended features, CNN+RNN (recurrent neural network) and ResNet.

For attention mechanism, its efficiency has been demonstrated in many fields such as speech recognition [14]. It makes the model focus on the specific features which are more relevant to the outputs. Therefore, it also attracts more and more researchers to improve their model performance in the field of SER. For example, Li et al. [15] proposed the combining use of the dilated residual network (DRN) and multi-head self-attention for SER task. Yoon et al. [16] proposed a multi-hop attention model to combine the features extracted from acoustic and textual data for SER task.

3. STSER Framework for SER

In this section, the proposed STSER framework combining speech and text information for SER is introduced. We firstly describe the overview structure of the proposed STSER. And then, we describe the designed smodel and tmodel. Finally, the multi-scale fusion strategy is presented.

3.1. Overview Structure of STSER

In this paper, for addressing the task of SER, we propose a multi-scale fusion framework based on the deep learning model by using both the speech and text data. The overview structure of the proposed STSER framework is shown in Figure 1.

Comparing with traditional methods using a large set of hand-crafted features (e.g., pitch, MFCC, etc.) produced from raw speech data, only the log-mel spectrograms obtained from speech data and text data are used as the input of the proposed STSER. It can save a lot of time and energy in designing and producing distinguished and suitable features, which always needs professional knowledge. In other words, it can eliminate the need to extract features manually.

As shown in Figure 1, the log-mel spectrogram and transcribed text data are processed by the smodel and tmodel, respectively. After extracting the features from two models, besides being treated separately, two feature vectors are concatenated into one vector and then fed into the corresponding classifier to further process. Finally, the emotion predictions are given by the ensemble of all classifiers.

3.2. Design of smodel

The smodel is used to learn speech representation from the input of mel-spectrogram for further classification. In our design of the smodel, we take advantage of CNN, Bi-LSTM and attention mechanism. The structure of the designed smodel is shown in Figure 2. There are mainly three components in the smodel, including CNN layers, Bi-LSTM layer and Attention layer.

3.2.1. CNN Layers

It has turned out that the CNN-based models can achieve a comparable or even better performance than traditional models using hand-crafted features. Therefore, we firstly use the CNN layers to extract robust features.

As shown in Figure 2, the CNN layers consist of one or more conv blocks. There are three types of layers in each conv block, including the convolutional layer, batch normalization layer and max-pooling layer.

- The convolutional layer is utilized to extract features and learn local correlations through the convolution operation of kernel filters. The kernel weights are shared to

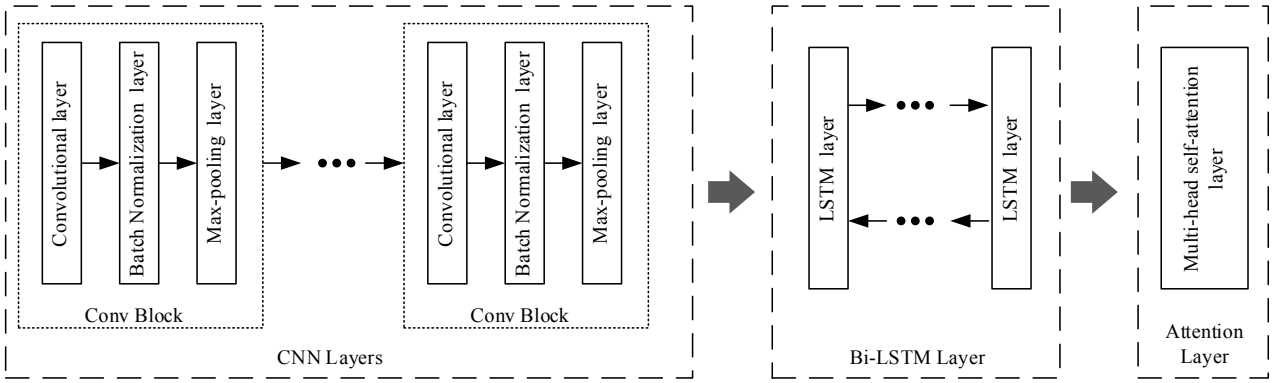

Figure 2: *Designed structure of the smodel*

extract the same features in different locations, which can greatly reduce the network parameters.

- The batch normalization layer is applied to stabilize or even accelerate the training process through normalizing the input distribution of this layer. By using the layer, it also can be able to alleviate the overfitting problem and then improve generalization performance.
- The max-pooling layer is used to reduce feature dimensions while retaining the corresponding main features. Besides, it also can prevent overfitting.

3.2.2. Bi-LSTM Layer

The main issue affecting the performance of the RNN, which is used for learning sequential information, is that it is easy to encounter the gradient disappearance problem. To mitigate this problem, LSTM is applied to learn long-term dependencies and contextual information by introducing the gating mechanism.

For the conventional LSTM network, it can only be able to predict the output of the next time step according to the sequential information of the previous time step. However, it does not consider that the output at the current time step is not only related to the former state, but may also be associated with the future state. Bi-LSTM, which uses two LSTM networks with forward and backward directions, can be applied to address the above situation. Considering the complexity of emotional changes in speech, we use Bi-LSTM layer to learn long-term dependencies of the extracted features from the CNN layers.

3.2.3. Attention Layer

As there are many short silence periods and only specific parts contribute to the emotion, it increases the difficulty and challenge of recognizing emotion. In order to address this challenge, the attention mechanism is introduced in the smodel.

Especially, multi-head self-attention mechanism [17] is utilized in the attention layer. With this attention layer, it is able to make the smodel focus on salient features related to the emotions and then generate discriminative representations for improving emotion recognition accuracy.

3.3. Design of tmodel

The tmodel is utilized to learn text representation from the transcribed text data as shown in Figure 1, for assisting the emotion classification. The ALBERT model [10] has been validated to be able to generate robust text feature representations with fewer model parameters compared with BERT [18] in NLP fields. Therefore, the text representation is directly obtained by using a pre-trained ALBERT model and then used for further emotion classification.

3.4. Multi-scale Fusion Strategy

For obtaining better classification performance, we present a multi-scale fusion strategy, including feature fusion and ensemble learning. Feature fusion is used to fuse extracted features, and ensemble learning is used to integrate multiple models. Both two approaches aim to improve the overall performance.

- For feature fusion, we generate a new representation vector by directly concatenating the speech and text representation, and this vector is then fed into the bimodal emotion classifier.
- For ensemble learning, three classifiers, including speech, text and bimodal emotion classifier, are used to give the emotion prediction separately. And then, the final predicted emotion is obtained by the average ensemble of all three classifiers in the paper.

4. Experiments

To evaluate the performance of the proposed framework, we perform experiments on the public emotion dataset - IEMOCAP. In this section, the experimental setup is firstly described. And then, the experimental results and analysis of the proposed framework are presented.

4.1. Experimental Setup

4.1.1. Emotion Dataset

There are approximately 12 hours of speech data from 10 actors in the IEMOCAP dataset. The dataset contains 5 sessions and each session is performed by one female and male actor in scripted and improvised scenarios. It consists of 10039 utterances in total, with a mean duration of 4.5 seconds. In this paper, the sample rate is set to 16 kHz. Besides, the speech utterances labeled 'excited' are merged into the 'happy' class, and then the data labeled $\{happy,\ neutral,\ angry,\ sad\}$ are used in the experiments. The IEMOCAP dataset contains the transcribed text data of each utterance with word-level alignment; hence, log-mel spectrograms obtained from speech and the corresponding transcript are used as the input of the STSER.

Table 1: *Performance comparison of each classifier with STSER framework in terms of WA and UA on IEMOCAP dataset.*

Method	WA	UA
Speech Emotion Classifier	53.35%	55%
Text Emotion Classifier	59.32%	59.4%
Bimodal Emotion Classifier	70.83%	71.7%
STSER	71.06%	72.05%

4.1.2. Hyper-parameters

Because the low dimensional log-mel spectrogram is processed to obtain speech representation, we only use one conv block in the CNN layers of smodel in the experiments. All three classifiers have three dense layers with the same hidden units, i.e., 128, 64, 128. There are four parallel heads in the attention layer. Adam optimization algorithm with a learning rate of 10^{-4} is applied to optimize the model parameters. Besides, the rectified linear unit (ReLU) function is used as the activation function of the layers in the model, including convolutional layers, Bi-LSTM layer and dense layers.

4.2. Experimental Results and Analysis

In our experiments, the STSER framework is trained in a speaker-independent manner. The five-fold cross-validation is implemented for evaluating the performance of the proposed STSER framework. In other words, the speech and text data of eight speakers from four sessions are used as training dataset for training the model, and the remaining data are used as testing dataset. In the training procedure, 20% of the training data are used as the validation dataset for further preventing overfitting and improving generalization abilities by applying the early stopping mechanism according to the validation accuracy.

With respect to the evaluation metrics for measuring the model performance, weighted accuracy (WA) and unweighted accuracy (UA) are employed in the experiments. Especially, WA is defined as the overall accuracy of the entire testing dataset, while UA is the average of accuracies of all different emotion categories. It is worth noting that the WA and UA reported in the paper are the average of the results of five-fold cross-validation experiments.

4.2.1. Ablation Study

As depicted in Figure 1, there are three classifiers in the STSER framework, i.e., speech, text and bimodal emotion classifiers. We compare the performance of three classifiers and the STSER framework.

As shown in the Table 1, the STSER framework performs the best accuracy of emotion classification among all models. However, the results of STSER framework and bimodal emotion classifier are close, it may be due to the poor performance of other two classifiers, resulting in the average result that could not greatly improve performance. In addition, the performance of bimodal emotion classifier is significantly better than both speech and text emotion classifiers. It represents that the feature fusion strategy can make full use of the multi-modal features and achieve complementarity. These two experimental phenomena indicate that the effectiveness of the multi-scale strategy, i.e., feature fusion and ensemble learning, for improving emotion recognition.

Table 2: *Performance comparison of the STSER framework with other methods using both speech and text data.*

	Method	WA	UA
[19]	E_vec-MCNN-LSTM	64.9%	65.9%
[20]	MDRE	71.8%	-
[21]	Att-align	72.5%	70.9%
[22]	IIEmoNet(BE)	73.5%	71%
Ours	STSER	71.06%	72.05%

4.2.2. Comparison Experiments

Experiments are performed to compare the proposed STSER framework with other SER approaches on IEMOCAP dataset. Table 2 shows the performance comparison of all methods in terms of WA and UA.

As shown in the Table 2, on one hand, compared with [21] and [22], the absolute accuracy of the STSER framework is reduced by 1.44% and 2.44% in terms of WA, respectively. This reduction is so small that it can be assumed that the STSER has achieved a comparable performance. On the other hand, the inputs of these models, especially the speech features, differ significantly. Only the log-mel spectrogram is used as the speech features in the STSER; however, in [21], it needs to extract a 34-dimensional feature vector from each frame including MFCC, spectral centroid and so on; in [22], the authors train their models on the IS09 feature set [23]. It indicates that more handcrafted features are needed to achieve a better performance. Nevertheless, more necessary feature inputs mean that more time and energy consumption in the preprocessing original speech data. It even needs more extra professional knowledge in this process, which further increases the complexity and difficulty of the SER task. Therefore, our proposed STSER framework may be a better choice as balancing the tradeoff between the model complexity and recognition accuracy. In addition, there are multiple classifiers in the STSER framework, which can be applied individually or in combination. It also has great extensibility for further integrating more classifiers about other modal information.

5. Conclusions

In this paper, we present a multi-scale fusion framework named STSER by combining speech and text data for addressing the SER task. Speech data is processed by smodel, which takes advantage of the CNN, Bi-LSTM and attention mechanism. Especially, CNN layers are used to learn local correlation, Bi-LSTM layer is applied to learn long-term dependencies and contextual information, and the multi-head self-attention layer is adopted to learn the salient features contribute to emotions and ignore silence and unimportant parts. A pre-trained ALBERT model named tmodel in the paper is used to process text data to generate text representation. Finally, a multi-scale fuse strategy, including feature fusion and ensemble learning, is applied to obtain a better and more accurate emotional prediction. The proposed model is evaluated on the public emotion dataset - IEMOCAP. The experimental results demonstrate that our model can achieve comparable recognition accuracy for SER with fewer feature inputs. In the future work, we will evaluate the performance using other network structures, such as residual connection, dilated network and so on. Furthermore, we will try to learn discriminative features and robust representations by using more modal information.

6. References

[1] M. M. H. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[2] J. Li, L. Lu, C. Liu, and Y. Gong, "Improving layer trajectory LSTM with future context frames," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 6550–6554.

[3] M. Tahon and L. Devillers, "Towards a small set of robust acoustic features for emotion recognition: Challenges," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 24, no. 1, pp. 16–28, 2016.

[4] S. Demircan and H. Kahramanli, "Feature extraction from speech data for emotion recognition," *Journal of Advances in Computer Networks*, vol. 2, pp. 28–30, 01 2014.

[5] J. Kim and R. A. Saurous, "Emotion recognition from human speech using temporal information and deep learning," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association*, 2018, pp. 937–940.

[6] W. Chen, D. Xie, Y. Zhang, and S. Pu, "All you need is a few shifts: Designing efficient convolutional neural networks for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7241–7250.

[7] W. Zhang, X. Cui, U. Finkler, B. Kingsbury, G. Saon, D. S. Kung, and M. Picheny, "Distributed deep learning strategies for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 5706–5710.

[8] Y. Hifny and A. Ali, "Efficient arabic emotion recognition using deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 6710–6714.

[9] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, H. Meier, and B. W. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2011, pp. 5688–5691.

[10] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," *CoRR*, vol. abs/1909.11942, 2019.

[11] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[12] M. Wöllmer, F. Eyben, S. Reiter, B. W. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association*, 2008, pp. 597–600.

[13] D. Tang, J. Zeng, and M. Li, "An end-to-end deep learning framework for speech emotion recognition of atypical individuals," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association*, 2018, pp. 162–166.

[14] N. Moritz, T. Hori, and J. L. Roux, "Triggered attention for end-to-end speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 5666–5670.

[15] R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, "Dilated residual network with multi-head self-attention for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 6675–6679.

[16] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 2822–2826.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, 2017, pp. 5998–6008.

[18] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[19] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, B. Yegnanarayana, Ed. ISCA, 2018, pp. 247–251.

[20] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE Spoken Language Technology Workshop, SLT 2018, Athens, Greece, December 18-21, 2018*. IEEE, 2018, pp. 112–118.

[21] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 3569–3573.

[22] V. Heusser, N. Freymuth, S. Constantin, and A. Waibel, "Bimodal speech emotion recognition using pre-trained language models," *CoRR*, vol. abs/1912.02610, 2019.

[23] B. W. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*, 2009, pp. 312–315.