Global Scientific JOURNALS
GSJ: Volume 8, Issue 5, May 2020, Online: ISSN 2320-9186
www.globalscientificjournal.com

ASVP-ESD: A dataset and its benchmark for emotion recognition using both speech and non-speech utterances

Dejoli Tientcheu Touko Landry, Qianhua He, Haikang Yan, Yanxiong Li*

School of Electronic and Information Engineering, South China University of Technology, Guangzhou, 510640, China

201722800077@mail.scut.edu.cn, eeqhhe@scut.edu.cn, haikangyan@163.com, eeyxli@scut.edu.cn

Abstract

In the field of human-computer interaction, speech emotion recognition has become a very challenging topic in the past decade. Emotion recognition has made great progress with the advancement of deep learning. It is worth noting that non-speech also contains rich emotions, current research mainly focused on the effectiveness of emotional features of the speech, implementing on public data set where basically most are speech based data set instead of considering non-speech emotional data. To deal with this challenge, this paper presents a realistic data set called ASVP-ESD(Audio,Speech and Vision Processing Lab Emotional Sound database), which contains six categories of emotions in speech and non-speech respectively, including 5105 samples with about 5 seconds length average collected from online videos. Several approach that achieved a STOA results on IEMOCAP and Berlin EmoDB data sets were implemented as reference for the new challenge. In order to improve the recognition performance, a method based on Convolution Neural Network- Bidirectional Long Short Term Memory(CNN-BLSTM) was proposed as the benchmark of this new data set achieving a recognition accuracy of 69.68%, which is 2% more higher than published method. Meanwhile a two-stage recognition strategy for this ASVP-ESD recognition task were proposed, as it can improve the recognition accuracy rate to 5%∼6% base on above published method.

Index Terms: speech and non speech emotion recognition data set, deep neural network, two-stage recognition strategy

1. Introduction

Emotion recognition is a process of identifying human emotion. In our daily life, emotion can be recognized through basic modality such as facial, speech, body gesture and text. Emotion recognition system may be used in an in-car driving system, where information about the mental state of the driver shall be provided and used to keep him alert during driving[1]. Emotion analysis of telephone conversation between criminals could help the police in their investigation. In medical field doctor can diagnose mental disorders through analyzing emotional contents of a patient's speech[2]. One of the major problems faced today in speech emotion recognition (SER) is that there are many kinds of emotions and some have similarity, obtaining these speech samples under realistic conditions constituted a very challenging problem. Numerous studies put more effort on recognition model architecture and corresponding features, even-though there are no specific features in SER. In the study of SER, one of the major challenge is the limitations of realistic data set, as most of the data sets are speech based, scripted in a clean scenario, where emotion could be exagerated than natural. The performance of Automatic emotion recognition system developed by such data set will significantly decrease when applied in real life. Therefore, ASVP-ESD was proposed as the new benchmark data set in speech and non-speech emotion recognition task. ASVP-ESD is a realistic and more natural emotional corpus collected from movies, Youtube channels, some utterances were recorded from real life human interaction in natural environment with no language restriction. The data set contains 5146 samples, 60% are non-speech emotional sounds and 40% are speech utterances, where 53% of samples are from female and the rest from male.

In order to improve efficiently the performance of emotion recognition considering speech and non- speech utterances based on ASVP-ESD database, experiments were done focusing on 6 different basic categories of emotion such as happy, angry, neutral, fear, sad and surprise. To evaluate the new challenge of ASVP-ESD, one state of the art published approache and the proposed CNN-BLSTM in this paper are implemented and evaluated fairly on ASVP-ESD. The proposed two-stage recognition strategy is described in Figure 1. Using the proposed two-stage recognition strategy, the recognition accuracy rate of 74.39% was achieved with the proposed CNN-BLSTM model which is a bit higher than 71.11% accuracy obtained by using the published Zhao et al[4] model. Meanwhile, the two-stage strategy is evaluated with different published approaches and get 5%∼6% improved performance than single stage in every approach. On the other hand, the proposed CNN-BLSTM method achieved an average accuracy of 83.71% on Berlin database and a state of the art(STOA) performance of 66.7% on IEMOCAP database which is fair comparing with the accuracy of 95.89% and 66.3% obtained by Zhao et al[4] and Yeh et al[8] on Berlin Emodb and IEOMCAP respectively.

In summary the main contributions of this paper are two folds: First, the collected ASVP-ESD emotional database, is more realistic and containing speech and non-speech utterances. Second, a two-stage recognition strategy architecture that achieved much better recognition result than using a single model. The rest of this paper is organized as follows. In section 2, related work and recent performance of deep learning in SER. Description of the emotional database used for this work in Section 3, then describe the details of the proposed methodology in Section 4, followed by experimental results in section 5. The final section is about our conclusion in Section 6.

2. Related work

Earlier studies in SER dealt only with whether the speech is positive or negative. Speech plays an important role in the society, it can contains information about the speaker, language, emotion state and many others. Speech has been proven to be among promising modality for recognition of human emotions.

*Corresponding author: eeqhhe@scut.edu.cn;

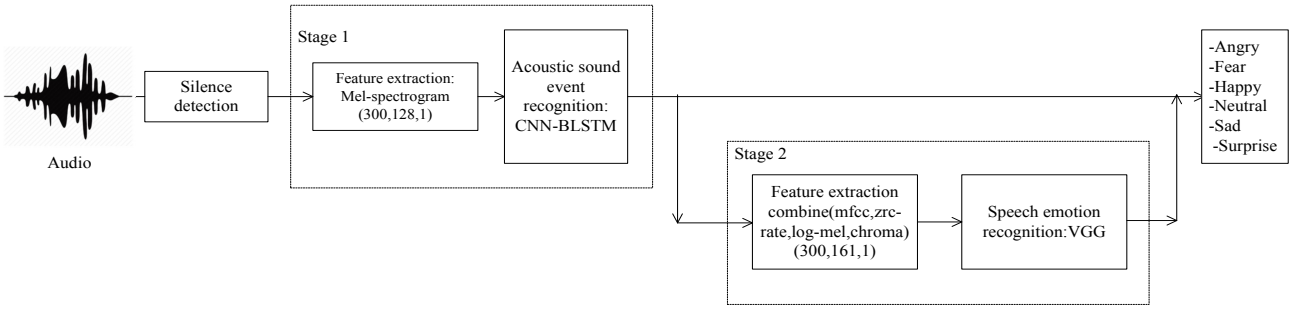

Figure 1: *Illustration of the two-stage strategy model architecture process for speech and non-speech emotion recognition.*

Many steps towards speech emotion recognition have already been taken by researchers with significant improvement. In the past years, several works have been done in SER using different methods including traditional classifier such as Support vector machine[1], Hidden Markov Model and Gaussian Mixture Model[3]. With the rapid development of deep learning, methods such as Convolution neural network(CNN)[4], combination of Convolution neural network with Recurrent neural network[4], Attention mechanism[5], Attention based convolution recurrent neural network[6, 7, 8] have demonstrated their effectiveness. In SER field while mutiple study and research achieved acceptable performance using different approach and methods, most of the state of the art results were obtained using Convolution based methods. Zhao et al[4] introduced two similar model architecture designed by the combination of convolution neural network and long short term memory(CNNLSTM) to learn local and global emotion related features from speech. Both network got a considerable result on two public emotional databases, achieving a recognition accuracy rate of 52.14% on IEMOCAP database[9], and a state of art of 95.89% accuracy on Berlin Emotional database[10]. Recently research brought some contribution in SER by implementing attention mechanism. Chen et al.[6] used a 3-D Convolution Recurrent Neural Networks with Attention mechanism Model to learn discriminative features, assuming that compared to 2-D convolution 3-D convolution can better capture more effective information for SER. Even-though the above studies have been successfully applied in SER field, they mainly focus on the effectiveness of the extracted features and the architecture of the model to achieved good recognition performance of emotion state. None of the work mentioned above have taken into consideration non-speech emotions.

In daily communication, human can also use emotional voice sound interjection to express their emotional state such as laugh for happy, cry for sad. Non-speech events can carry rich emotion information that can be helpful to clearly determine the speaker emotion. Among the few works done by taking non-speech into consideration, Huang et al[11] proposed a method using deep neural network considering verbal and nonverbal segments, using the NNIME emotion corpus utilizing 9384 verbal segments and 5252 nonverbal segments such as laugh(183), breath(409), shout(67), silence(4593)[12]. The first process consisted of an SVM-based verbal/nonverbal sound detector application, then an auto-tagger was employed to extract the verbal/nonverbal segments, emotion and sound features were respectively extracted based on convolution neural networks and then concatenate to form a feature vector used as input to a sequence to sequence long short term memory based model to output an emotional sequence as recognition result, and a detection accuracy of 52% was achieved. We can observe that the CNNs based methods used to extract generic features, can be helpful to learn local and global features from the sound/speech segment. The model architecture complexity can affect the final result, also non verbal segments data are limited and unbalance which is a big challenge for deep learning data-driven. Meanwhile, there are some well-known emotional speech based data sets , such as IEMOCAP, Berlin EmoDB.

In SER field, there are few emotion data sets that contains both speech and non-speech emotional utterances. Compared to non-speech utterances, speech beside containing emotion also have rich and semantic content. This extra semantic content leads to a greater distinction between speech and non-speech utterances itself. Therefore we proposed a two-stage recognition strategy considering sound features information of speech and non-speech in the emotion recognition process, to improve the recognition performance.

3. Data set description

In this study, ASVP-ESD emotional database[1] was used as benchmark emotional database. It is a human voice emotion corpus containing speech and non-speech sounds. This database differ from other scripted public emotion speech based database. The database is among the few emotional data sets that contains a variety of emotional non-speech sound such as laughing, crying, screaming, rage and amazed. Speech data contains angry, happy, neutral, fear, sad and surprise emotional state.

3.1. Acquisition process

Voice sounds and utterances of the database were collected in movies, youtube channel and from multiple online human emotion voice sound website. The sample were collected without language restriction. The database only contains audio sample in wave format with 1 channel and sampled rate of 16k, audio length average are 5 seconds, the total duration is 6.6 hours, it is a 1.04G data set.

3.2. Labeling process

The Database labeling process was done by 5 different annotators through a tagging application specially design for audio tagging. After listening to each audio the judge choose the corresponding label according to their personal feeling. When the tagging part was done, a simple voting algorithm was build for voting and upgrading the corresponding audio to the class hav-

[1] https://zenodo.org/record/3782416 Free download of the data set

Table 1: *ASVP-ESD emotional database description used in this work*

Non-speech	speech
Rage (18)	Anger (602)
Cry (892)	Sad (448)
augh (772)	Happy (302)
Neutral (249)	Neutral (453)
Scream (679)	Fear (62)
Amazed (494)	Surprise (134)
Total: 3104	Total: 2001

ing the most number of vote. When equality occur the emotion was randomly choosed between the class with equal vote.

3.3. Data organization

ASVP-ESD data set contains a total of 5105 audio sample, which is arranged in 55 folders, where 26 folders are for male voice samples and 29 folders are for female voice. Table 1 shows the distribution of the commonly used emotion types in ASVP-ESD. In this work 4940 samples were used as audio with less than 10 frames were not considered.

4. Approach

Speech emotion recognition contains two very essential modules such as : Feature extraction and classification. The process used in this research are similar, Figure 1 shows the proposed process framework. Emotion can be clearly and more precisely differentiated through non-speech sound, where in speech the volume and tonality of the sound within a specific duration are important factors for determining the corresponding emotional state. Both speech and non-speech sounds have a strong relationship to each other and are complementary. In first stage training process, silence detection was applied to obtain speech and non-speech emotional sound segments. Features were extracted from sound segments then used as input of the model first stage recognition.The input used on this network architecture is a 128 mel-scale spectrogram and 161 frequencies bins for combined features(mfcc, log-mel spectrogram, zero -crossing rate and chroma) in the first and second stage respectively. Their are personalized features that achieved great result in SER due to their ability to carry speaker personal emotional information.

4.1. Model architecture

The proposed CNN-BLSTM model was used in the first stage to extract acoustic information from the inputs. BLSTM have the ability to better understand the context, preserve the information form the past and future by running the inputs in two ways. The model convolution and pooling are two dimensional, the number of convolution filters was increased by factor of two as the network get deeper as shown in Figure 2 and table 2. Small size(3x3) kernel was applied to extract small complex sound features on a smaller receptive field of the mel- spectrogram. This architecture was suitable to capture neighborhood features in a small scale, as non-speech emotional sounds can occur in a limited short duration. Max-pooling was used to provide more statistical information for the following layer. This model was constructed to capture and learn high-level emotional features from the input mel spectrogram. The last layer was used to map these features into the required output-space. The model achieved great perfromance in sounds classification task due to

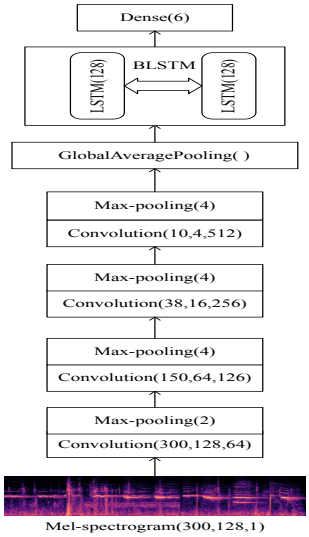

Figure 2: *Bidirectional Long short term memory .*

Table 2: *The layer parameter of the CNN-BLSTM network, with output dimension represented as height x width x number of filter. M x N is the size of the features, the last layer kernel size k is the number of the emotional events.*

Name	Output Dim	kernel
con2-1	MxNx64	3x3
Max-pooling2-1	M/2xN/2x64	2x2
con2-2	M/2xN/2x128	3x3
Max-pooling2-2	M/8xN/8x128	4x4
con2-3	M/8xN/8x256	3x3
Max-pooling2-3	M/32xN/32x256	4x4
con2-4	M/32xN/32x512	3x3
Max-pooling2-4	M/128xN/128x512	4x4
GlobalAveragePooling	-	-
Bidirectional-LSTM	-	128
Dense	-	k=6

his flexibility.

VGG(CNN-VGG16) model architecture is similar to other classification models, with 7 convolution layers originated from VGG[13] performed on Task 1 to 5 of DCASE 2018. The advantage of using VGG network is that it exploit timely-local correlation by enforcing a local connectivity pattern between the input and the CNN neurons[14]. CNN are able to learn relevant features from specch sound signal at different levels similar to a human brain. In this model Blocks containing two sequential convolution layers with the same filter size applied multiple times to extract more complex and representative features. Max-pooling was preferred because it generalize better the result from a convolution filter. The last three layers are dense layers with 256, 128, and 6 hidden units respectively, which are used as classifier.

For both model Global Average pooling was applied to improve the model performance and be helpful to reduce overfitting because there is no parameter to be learned in that layer, Softmax function was applied to the last Dense layer. Elu(Exponential Linear Unit) are used as the activation function in all convolution layers instead of Relu(Rectified linear units)

Table 3: *Model performance on public dataset based on speaker independent. Evaluate with average accuracy(A.A) and unweigthed average recognition(UAR)*

Research	Berlin(A.A)	IEMOCAP(UAR)
Yeh et al[8]	-	66.3
Chen et al[6]	82.82	64.74
Zhao et al[4]	95.89	52.14
Our CNN-BLSTM	83.71	66.7
State of the art	95.89	66.7

Table 4: *Recognition accuracy comparison on ASVP-ESD (combine speech & non speech) single stage and two stages.*

stage one	stage two	Accuracy%
VGG	-	65
Zhao et al[4]	-	67.42
Our CNN-BLSTM	-	69.68
Zhao et al[4]	Zhao et al[4]	71.11
Zhao et al[4]	VGG	72.04
Our CNN-BLSTM	VGG	74.39

because it tend to lead the function to converge faster and produce more accurate results. Our proposed CNN-BLSTM can extract high level features in emotional sound event recognition task, where VGG can capture information more effectively for the speech emotion recognition task.

5. Experiments and results

This section present the experimental results used to evaluate the approach in section 4.

5.1. Implementation details

Model architectures were built with Keras. Sound signal was splited into segment with length of 300 frames as speaker emotion can be detected in a short period of time, to also make the process smooth. In the training process each segment was taking as training sample, while during the testing phase the result was obtained through a voting process after evaluating the whole sample prediction. The mean and standard deviation was calculated for the network input. An Adam optimizer[15] with learning rate of 0.001 was applied, loss function was categorical cross entropy for all the experiments. The batch size are 32 samples with 500 epochs. L2 regularization of 0.005 was applied to prevent over-fitting. Global average pooling was used for feature map on the output of the local convolution layer with dropout[16] of 0.5.

5.2. Evaluation and comparison with baseline

In this evaluation, experiments were implemented using Zhao et al[4] recent research model structure. This model have good effect on public speech based data set as shown in table 3, moreover beside achieved high emotion recognition accuracy, the model also have better generalization ability. CNN and LSTM are combined together to learn the high-level features, which contain both the local information and the long-term contextual dependencies[4].The model contains four local feature learning block(LFLB), with Each composed of convolution, Batch-Normalization which improves the training of the model significantly, activation and max-pooling layer. The network width are 64 and 128 respectively for LFLB1, 2 and LFB3, 4. The training results are relatively stable, but different calculation between training and testing was observed as the model converge better and fast on the training set.

5.2.1. Evaluation on ASVP-ESD database using a single stage

In this experiment speech and non-speech were combined together, based on six emotions (happy with laugh, sad with cry, fear with scream, surprise with amazed, neutral and angry)Beside using Zhao et al[4] model, experiments was also carried on VGG in a single stage sturcture similar to traditional architecture. As shown in table 4 VGG model achieved a recognition accuracy of 65%, where Zhao et al[4] model achieved a recognition rate of 67.42%, 2.42% higher than the VGG but lower than our proposed CNN-BLSTM model that achieved a performance of 69.68%. In this experiment, Zhao et al[4] model effect decreases significantly when applied on ASVP-ESD database, due to the fact that the database is more realistic different from other scripted public data sets, it was recorded in normal environment condition with high present of noise in some samples.

5.2.2. Evaluation on ASVP-ESD database using a two- stage strategy

Our proposed network width starts at a small value of 64 and increases by factor of two after every sub-sampling/pooling layer, more filters as the network gets deeper can be an advantage for better convergence. In this experiments the first stage consisted of recognition of six different emotional sounds event considering speech uttereances as an emotional event. We adopted a 4 folder-cross validation on the data set, for each experiment one folder was used as validation data set and the 3 others as training data set, the proposed model achieved an average accuracy of 82.7%. Table 5 shows the confusion matrix of the corresponding stage result. As speech signals are time-varying signals that need special processing to reflect time varying properties, the second stage consisted of emotions recognition through speech focusing only on utterances classified as speech sounds in the first stage, which were used as validation set in this stage where Speech sounds data from others corresponding training folder were considered as training data set. The VGG model performe better on this stage with an average recognition rate of 63.4%. Table 6 shows the confusion matrix result of the second stage. Table 7 and table 8 showed evaluation metrics and confusion matrix of the final result output obtained by concatenating both stage results, where laugh stand happy, cry stand for sad, scream for fear, amazed for surprise, neutral for neutral sound and angry. Mel spectrogram shows to be effective for the acoustic emotional recognition task where the combination of mfcc, log-mel spectrogram, zero-crossing rate and chroma are effective for the speech emotion recognition task. Zhao et al[4] model plus VGG achieved a recognition rate of 72.04%, higher than using the same model in both stages but lower than the combination made of our proposed method plus VGG that achieved a performance of 74.39% accuracy .

5.2.3. Evaluation on public data set

Experiements were also performed on Berlin Emodb and IEOMCAP data sets to evaluate the performance of our proposed method on public data set. IEOMCAP contains a total of 10039 utterances with10 speakers(male and female) sepa-

Table 5: *First stage CNN-BLSTM Confusion matrix of emotional sound event classification on 1277 testing sample*

	Neu	**Lau**	**Cry**	**Scr**	**Ama**	**Sep**
neutral	23	3	1	14	15	7
laugh	3	157	3	5	2	23
cry	3	13	180	7	5	15
scream	2	2	7	149	4	8
amazed	8	5	8	3	92	9
Speech	1	4	14	1	2	479

Table 6: *Second stage VGG classifier focus on sample classed as speech sound in first stage*

	Neu	**Hap**	**Sad**	**Ang**	**Fea**	**Sup**
neutral	69	11	18	19	1	1
happy	12	61	4	14	1	0
sad	18	10	77	6	2	0
angry	16	9	9	115	1	0
fearful	4	3	6	6	2	0
surprise	19	6	5	7	1	8

rated in 5 sessions. Berlin Emodb contains 535 utterances display by 10 actors(male and female) in seven different emotions categories. In this evaluation, for IEMOCAP we considered four emotions(angry, happy, sad and neutral), sessions 1 to 4 were usied as training data and session 5 as validation data, for Berlin EmoDB we randomly take 2 speakers(1 male ,1 female) for validation data and the 8 others as training data. The best features used for IEOMCAP was mel-spectrogram where for Berlin EmoDB we used the combined features vector to obtain better performance. Our model achieved an average performance of 83.7% on Berlin data set and a STOA unweighted accuracy recognition of 66.7% on IEOMCAP, outperforming the 66.3% obtained by Yeh et al[8].

5.3. Experiment Analysis

The best recognition accuracy for a single stage model when combined speech and non-speech utterances was 69.68% as mentioned in table 4. Our proposed two-stage strategy achieved much higher accuracy regardless of which method is used comparing to single stage. Table 4 also shows the advantage and effectiveness of our framework process by first proceeding to the classification of different emotional sounds, as each non-speech emotional sound refer to a specific emotion category. Neutral class have the lowest recall value due to the fact that, the class has limited sample and high level variation between

Table 7: *Final evaluation metrics,using precision(P), recall(R) and f1-score(F1)*

	P	**R**	**F1**
neutral	0.52	0.53	0.52
happy	0.78	0.83	0.80
sad	0.81	0.81	0.81
angry	0.69	0.76	0.72
fearful	0.81	0.82	0.81
surprise	0.79	0.61	0.69
Accuracy			0.74

Table 8: *Final confusion matrix*

	Neu	**Hap**	**Sad**	**Ang**	**Fea**	**Sup**
Neutral	92	14	19	19	15	16
Happy	15	221	9	14	6	2
Sad	21	23	269	6	9	5
Angry	16	10	9	115	1	0
Fearful	6	5	13	6	151	4
Surprise	28	11	13	7	5	102

neutral sound including silence, yawn and other similar sound. In the second stage more surprise class sample were classified as neutral, listening to some sample we realized that few emotions with tight similarity occur in a same utterances example of surprise, realization, neutral, contempt there is a confusion among such utterances that makes the recognition less effective.

6. Conclusions

In this paper, due to the implementation of different network architecture used for recognition, observation can be done that CNN based model have an effective impact for emotion classification. The proposed two-stage strategy architecture have made 5%∼6% improvement on the recognition performance accuracy, where our proposed CNN-BLSTM model achieved fair recognition performance on public data set, showing its robustness for emotion recognition task. The proposed method can achieve better performance than others, specially on large data sets. In future work we plan to expand the number of emotional category, to use classes with similarity such as boredom and disgust, and also investigate the use of specify emotion categories in different level such as anger which is a primary emotion, it can be specify as irritation, exasperation, rage which are secondary emotion.

7. Acknowledgements

This work was supported by China Disabled Persons' Federation Fund CJFJRRB14-2019, and the National Nature Science Foundation of China (Grant No. 61571192, 61771200).

8. References

[1] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. I–577.

[2] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829–837, 2000.

[3] J. Kim and R. A. Saurous, "Emotion recognition from human speech using temporal information and deep learning." in *Interspeech*, 2018, pp. 937–940.

[4] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.

[5] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-attention for speech emotion recognition," *Proc. Interspeech 2019*, pp. 2578–2582, 2019.

[6] M. Chen, X. He, J. Yang, and H. Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[7] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech emotion classification using attention-based lstm," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1675–1685, 2019.

[8] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "An interaction-aware attention network for speech emotion recognition in spoken dialogs," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6685–6689.

[9] C. Busso, M. Bulut, C.-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, p. 335, 2008.

[10] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[11] K.-Y. Huang, C.-H. Wu, Q.-B. Hong, M.-H. Su, and Y.-H. Chen, "Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5866–5870.

[12] H.-C. Chou, W.-C. Lin, L.-C. Chang, C.-C. Li, H.-P. Ma, and C.-C. Lee, "Nnime: The nthu-ntua chinese interactive multimodal emotion corpus," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 292–298.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[14] W. Zhang, W. Lei, X. Xu, and X. Xing, "Improved music genre classification with convolutional neural networks." in *INTERSPEECH*, 2016, pp. 3304–3308.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.