Speech Emotion Recognition Using 1D CNN with No Attention

Yulan Li*, Charlesetta Baidoo*, Ting Cai*, Goodlet A. Kusi†
*School of Information and Software Engineering
†School of Computer Science and Engineering
University of Electronic Science and Technology of China
No.2006, Xiyuan Ave, West Hi-Tech Zone, 611731
Chengdu, Sichuan, P.R. China
gliyulan@gmail.com, charlsy18@outlook.com, ct_sxu@163.com, gkusikat@gmail.com

***Abstract*—Speech emotion recognition (SER) has gained much attention in recent years. SER system may be efficient depending on how much useful information contained in the extracted emotional features. Many research works have achieved state-of-the-art results using Convolutional Neural Network with different extracted speech features. These kinds of models can't collect relative emotional salient features from speech signal. In this paper, we present a novel complementary feature extraction method to extract salient emotional features. We compute Melspectrogram and Mel Frequency Cepstral Coefficients (MFCC) to capture time-frequency domain information, aimed at converting raw speech into emotional informative features from speech signals. Moreover, we adopt complementary property strategy to extract features and construct 1D CNN model which selects emotional features effectively and evaluate the model's performance on IEMOCAP, RAVDESS and Emo-DB speech corpus. Our method achieves better performance than baselines and competitive results using complementary features as input.**

***Index Terms*—Speech emotion recognition, 1D CNN, Complementary features, Speech emotion corpus**

I. Introduction

In recent years, application of machine learning has increased rapidly with the use of deep learning techniques in solving several recognition problems. Speech is a common and easy mode of communication amongst humans which enables people to interact with each other. During this interaction, they recognise emotional changes within a time frame based on their speech which makes emotion recognition an important area as far as Human Machine Interaction (HMI) is concerned. Other areas where speech emotion recognition can be applied include medical science, robotics engineering, automobile systems, calling centers, etc [1, 2]. Recognizing feelings from speech may be a very challenging issue to analyse since individuals express feelings in completely different ways.

In elicited emotional speech databases, there is the creation of artificial emotion which to some extent may not be available and the emotion expressed may not be original, such as actor based speech database which is collected from artists. It can provide real emotion when analysed and tested though. Natural speech database is very useful and suitable for the real world [3]. Thus, research in SER requires speech corpus obtained from an environment without noise to make results more accurate.

Speech emotion features may be categorised into spectrum features, prosodic features, non-linear features, and other features [4]. Some researchers have researched spectrum features in SER, which exhibits unique and valuable information, such as Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC), Log Frequency Power Coefficients (LFPC), etc. A typical example is the spectrogram as the inputs of a model as demonstrated in other CNN architectures [5].

Many researchers used different feature extraction methods such as MFCC as the main emotional feature or combined several methods to build deep learning models [6, 7]. Deep learning methods in SER can automatically accomplish feature extraction and can avoid the complex hand-crafted feature program. Convolutional neural network (CNN) and recurrent networks (RNN) application in speech emotion recognition have gained a quantum leap. Chen et al. [8] proposed an RNN model extracting log-Mels feature and combining with attention to build a layer for SER. Mirsamadi et al. [9] constructed B-LSTM model and applied local attention on speech signal which can make emotion more salient. Liu et al. [10] applied global k-max pooling and CNN to build a model. In their findings, they tried different k parameters which gained different results for Weighted Average Recall (WAR), Macro Average Precision (MAP) and Unweighted Average Recall (UAR). Many methods have been designed to achieve efficiency in speech emotion recognition but there is a setback when it comes to speech feature extraction which calls for solutions and interventions. In most works there have not been a clear response as to which feature carries the immanent emotion and also the accurate and most effective result for emotion recognition.

This paper proposes an efficient method using convolutional neural network to capture emotional features through extracting complementary features. We mainly extract Mel-frequency cepstrum coefficients (MFCC), melspectrogram and log-melspectrogram (LogMels) as feature maps. The model adopts 1D convolutional neural networks which achieves superior and competitive results and is evaluated by using three emotion corpus which are Emo-DB [11], IEMOCAP [12], and RAVDESS [13].

978-1-7281-2544-2/19/$31.00 ©2019 IEEE

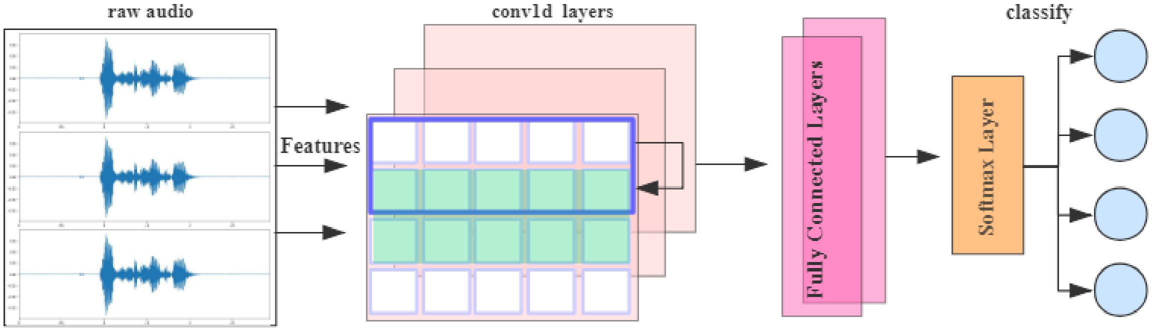

Fig. 1. The proposed None-attention 1D CNN (NACNN) architecture

II. Related Works

CNNs have gained significant results for recognition tasks spanning from image, video and sound. Wunarso and Soelistio [14], built a speech database (I-SpeED) based on the Indonesian native language. They extracted the amplitude, speech duration, and approximation coefficients as their inputs features using SVM. After their analysis and evaluation, they achieved an average accuracy of 76.84%.

Liu et al. [15] applied emotional feature fusion by combining spectral and prosodic features. They also used CNN's and DNN's to build their model. They further used Chinese emotion corpus (CASIA) which considered five emotions namely; angry, fear, happy, neutral, sad, and surprise. With the combination of spectral and prosodic features, their method's performance was improved in SER since their feature fusion fused a time and information domain together which has not been done in the past.

Guo et al. [16] proposed CNN-ELM; they built a model and used one feature fusion method, pre-processed speech signal by combining spectrogram features and heuristic-based discriminative features for their research. Their experiment results showed that feature fusion combined in CNN-ELM and the hybrid model was effective and more importantly they proved that heuristic features can make assignable contribution in SER system.

Deep learning methods have also been applied in SER. Jaybrata et al. [17] used MFCC method as emotion feature for input, they built a CNN-LSTM deep learning model. Mirsamadi et al. [9] used an attention mechanism and combined with BLSTM to design their network. They obtained salient emotional feature content and ignored opposite content, the results of the network achieved +5.7% and +3.1% in weighted accuracy and unweighted accuracy, respectively. Chen et al. [8] proposed a 3-D CRNN model and evaluated on Emo-DB and IEMOCAP corpus choosing only four emotions. Spectograms can be represented by speech signals which can be analysed by time characteristics. Puterka et al. [18] presented speech feature in time domain by applying different segments of speech signal which proved that the time domain of the spectrograms features can be efficiently applied to SER. The results also showed that between different times there can be a better accuracy. Lim et al. [19] also proposed the SER system based on CNNs and RNNs, adopted time-frequency analysis by transforming speech signal to 2D representation as inputs, used Emo-DB corpus and obtained time distributed CNNs results at an average precision of 92.02%, average recall of 88.42%, and average f-1 score of 89.56%.

Zamil et al. [20] proposed Mel Frequency Cepstrum Coefficient (MFCC) feature extraction from speech signals to identify the basic emotion of the speech. The extracted features were used to classify different emotions using LMT classifier. Using a voting mechanism on the classified frames, the emotion of the speech signal was detected. Two speech corpus; Emo-DB and RAVDESS were used to evaluate the model. Amongst the trained models, the best performance accuracy was 70% for seven different emotions.

III. Proposed Method

In this section, we introduce our proposed none attention method based on 1D CNN (NACNN) for SER and the method of feature extraction.

A. Complementary features

Speech of Melspectrogram (Mels) spectrogram and Mel-Frequency Cepstral Coefficients (MFCC) features lie in time-frequency domain. Melspectrogram and MFCC are extracted to capture time-frequency domain information, aiming to convert raw speech into emotional informative features from speech signals using librosa [21] to load audio files as a floating point time series. Furthermore, computation of LogMels (log Melspectrogram) to capture classes that lie in the different time-frequency domain is done. We conjecture that Mels and LogMels can capture complementary emotional information, so we combine the two complementary features (Mels and LogMels) in order to capture both higher frequency domain and lower frequency domain classes to get emotion information. we also calculate chromagram, spectral contrast [22] and the tonal centroid features [23] as feature inputs. The LogMels is calculated as follows:

$$l_i = log(1 + m_i) \tag{1}$$

Melspectrogram is extracted using librosa to produce output m_i, the LogMels l_i is produced by calculating the logarithm of m_i. When we load raw audios, the audios are either sampled to the original sample rate or resampled to the rate of 22050. We further extract MFCC features and the number of MFCCs is 40 as return and the length of the FFT (Fast Fourier Transform) window for extracting Melspectrogram is 2048 or 1024 to generate frequency spectrum, each frame of audio is windowed and the window is length of 2048 or 1024.

B. Model architecture

The model is constructed based on 1D CNN with fully-connected layer. Each 1D convolutional layer is followed by a batch normalization layer [24] to normalize the output in each layer in order to avoid varying distributions of the features across the training and test data. ReLU activation function is applied to contribute to the sparsity of the network which reduces the interdependence of parameters and alleviates the occurrence of over-fitting problem. As shown in Table I, the model has six 1D convolutional layers. After the 6^{th} 1D convolutional layer, the feature maps are flattened into a one column matrix in order to fit them into the fully connected layers. The flattened output is then fed to two dense layers and two dropout layers, then finally softmax function is applied for classification in the last layer. The feature map contains informative features as well as an emotional class label for utterance and loss calculation. In summary, the whole network is made up of six 1D convolutional layers and two dense layers with two dropout layers. The proposed model architecture summary is shown in Fig 1.

TABLE I
MODEL ARCHITECTURE

Layer	**Filter Shape**	**Output Shape**
1^{st} Conv1D	3 x 1 x 128	321 x 128
2^{nd} Conv1D	3 x 1 x 128	321 x 128
3^{rd} Conv1D	3 x 1 x 256	321 x 256
4^{th} Conv1D	3 x 1 x 256	321 x 256
5^{th} Conv1D	3 x 1 x 256	321 x 256
6^{th} Conv1D	3 x 1 x 256	321 x 256
Flatten	-	82176
1^{st} Dense	-	256
2^{nd} Dense	-	256
Softmax	Classifier	1 x 7

IV. EXPERIMENTS

In this section, experimental results for the three databases, Emo-DB, RAVDESS and IEMOCAP on NACNN architecture, are provided. Firstly, we experiment our baseline model **NACNN** without LogMels. Moreover, we also did experiment for LogMels as complementary features, here termed as **NACNN+LogMels**. Furthermore, we experiment NACNN for LogMels as complementary features and load raw audios adopting the original sample rate to extract features with setting the length of FFT window and each frame of audio windowed by the length of 1024 for extracting features for the model architecture, here by termed as **NACNN+LogMels***.

A. Databases and Results metrics

Emo-DB is an actor based emotional speech database consisting of 7 emotion classes (neutral, happy, angry, bored, fear, sad, and disgust) with 535 total samples [11]. The RAVDESS database is also an actor based speech corpus which is recorded by 24 professional actors (12 female, 12 male). RAVDESS contains 1440 samples in total consisting of calm, happy, sad, angry, fearful, surprise, and disgust expressions. IEMOCAP also contains 7380 samples which consists of angry, excited, frustrated, happy, neutral and sad expressions. For IEMOCAP emotion corpus, four emotion classes are selected (angry, sad, happy and neutral); the four emotion classes contained 1708 neutral samples, 595 happy samples, 1084 sad samples and 1103 angry samples. Happy had the smallest sample, hence, it is combined with that of the excited sample.

NACCN+LogMels was validated by computing the accuracy, precision, recall, f1-score and confusion matrix on the three databases. For a given test dataset, Accuracy is the ratio of the number of samples correctly classified by the classifier to the total number of samples. Accuracy can be defined as the following expression:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

TP (True Positive), FN (False Negative), FP (False Positive) and TN (True Negative). Precision is the positive samples are real positive samples, which are based on prediction results. Recall is positive examples in the sample have been predicted correctly, it is for original samples, also called recall rate. The weighted average of Precision and Recall can be simply described as F1-score. Precision, Recall and F1-score the mathematically expression as the following:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$F_\alpha = (1 + \alpha^2)\frac{Precision \times Recall}{\alpha^2 \times Precision + Recall} \tag{5}$$

$$F_1 = 2\frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

Weighted Average Recall (WAR) and Unweighted Average Recall (UAR) and Macro Average Precision (MAP) are computed to evaluate our model. MAP is one of the indicators to measure the performance of classifiers, which reflects the classification effect of classifiers on each category. The WAR, UAR, and MAP can be expressed mathematically as follows:

$$MAP = \frac{1}{S}\sum_{i=1}^{S}\frac{TP_i}{TP_i + FP_i} \tag{7}$$

$$WAR = \frac{\sum_{i=1}^{S} TP_i}{\sum_{i=1}^{S}(TP_i + FN_i))} \tag{8}$$

$$UAR = \frac{1}{S}\sum_{i=1}^{S}\frac{TP_i}{(TP_i + FN_i)} \tag{9}$$

Among them, TP_i is the correct number in the real case of category i, FP_i is the number in the real case of not category i, and S is the number of categories. FN_i is the failed number in the real category i.

B. Implementation details

The speech radios are mixed up for each database and 80% for training and 20% for testing is attained. Our network architecture based on Keras[1] using NVIDIA CUDA parallel computing platform and programming model on GPU. We trained the proposed model with a batch size of 32 across all databases using rmsprop optimizer with learning rate of 10^{-5} for 1000 epochs for the three databases. We adopt dropout layers and early stopping to prevent overfitting.

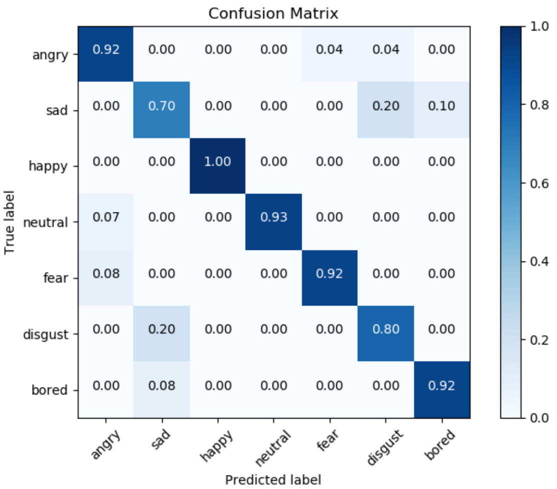

Fig. 2. Normalized Confusion matrix test accuracy for Emo-DB database for NACNN+LogMels

TABLE II
RESULT OF EXPERIMENT BASED ON EMO-DB FOR NACNN+LOGMELS

Emotion	Precision	Recall	f1-score	support
angry	0.92	0.92	0.92	26
sad	0.78	0.70	0.74	20
happy	1.00	1.00	1.00	5
neutral	1.00	0.93	0.96	14
fear	0.92	0.92	0.92	12
disgust	0.71	0.80	0.75	15
bored	0.86	0.92	0.89	13
avg/total	0.87	0.87	0.87	105

C. Results for Emo-DB database

NACNN+LogMels achieved good performance with an accuracy of 86.67% for the six emotions (neutral, anger, anxiety/fear, happiness, sadness, disgust and boredom) as shown in Table II. The total test samples were 105 for Emo-DB database. From the generated confusion matrix showed in Fig 2, happy shows the best recognition result; angry, neutral, fear and bored show better recognition result. In the normalized confusion matrix, sad is confused for disgust or bored, and disgust is confused for sad.

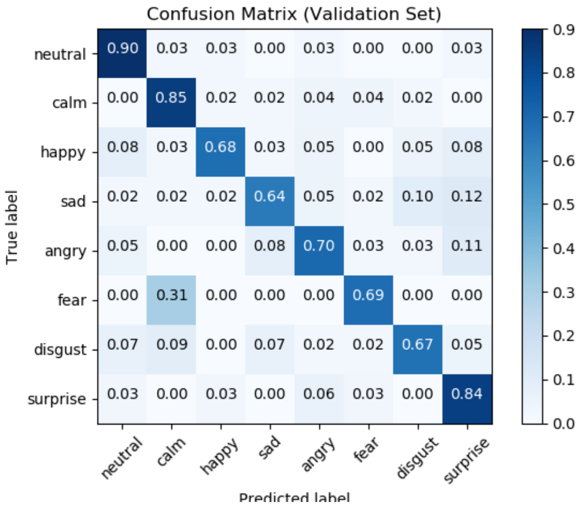

Fig. 3. Normalized Confusion matrix test accuracy for RAVDESS database for NACNN+LogMels

TABLE III
RESULT OF EXPERIMENT BASED ON RAVDESS FOR NACNN+LOGMELS

Emotion	Precision	Recall	f1-score	support
neutral	0.78	0.90	0.84	40
calm	0.77	0.85	0.81	47
happy	0.87	0.68	0.76	38
sad	0.77	0.64	0.70	42
angry	0.72	0.70	0.71	37
fear	0.65	0.69	0.67	16
disgust	0.78	0.67	0.72	43
surprise	0.64	0.84	0.73	32
avg/total	0.76	0.75	0.75	295

D. Results for RAVDESS database

NACNN+LogMels achieved accuracy up to 75.25% for the eight emotion (neutral, calm, happy, sad, angry, fear, disgust and surprise). As shown in Table III, we present results on precision, recall and f1-score for 295 test samples. From the generated confusion matrix shown in Fig 3, fear is often confused for calm whiles neutral, calm and surprise showed better performance.

E. Results for IEMOCAP database

NACNN+LogMels achieved average accuracy up to 65.35% for four emotion (neutral, happy, sad and angry). From Table IV, we show the results on precision, recall and f1-score

[1] Keras is a high-level neural networks API, written in Python and capable of running on top of TensorFlow - https://keras.io/

on about 1137 test samples. From the normalized confusion matrix from Fig 4, sad shows poor recognition of rate.

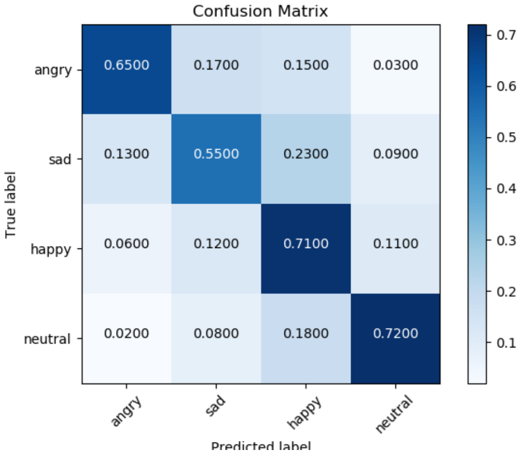

Fig. 4. Normalized Confusion matrix test accuracy for IEMOCAP database for NACNN+LogMels

TABLE IV
RESULT OF EXPERIMENT BASED ON IEMOCAP FOR NACNN+LOGMELS

Emotion	Precision	Recall	f-1 score	support
angry	0.69	0.65	0.67	229
sad	0.65	0.55	0.60	340
happy	0.63	0.71	0.67	365
neutral	0.66	0.72	0.69	203
avg/total	0.65	0.65	0.65	1137

F. Ablation studies

As shown in Table V, Emo-DB, IEMOCAP and RAVDESS datasets with none LogMels features, the **NACNN** method achieved accuracy 86.5%, 63.55% and 74.63% respectively and UAR 82.79%, 62.40% and 71.98% respectively.

The method **NACNN+LogMels** was loaded with LogMels as complementary features. This method achieved better accuracy than method **NACNN** for three datasets as shown in Table V.

For method **NACNN+LogMels***, we loaded raw audios and let audios be sampled to the original rate and the length of the FFT window for extracting Melspectrogram is 1024 to generate frequency spectrum, each frame of audio is windowed and the window is of length 1024. As shown in Table V, Emo-DB, IEMOCAP and RAVDESS datasets achieved accuracy 87.24%, 63.72% and 76.66% respectively, UAR 84.17%, 62.62% and 74.49% respectively.

The results for **NACNN** showed the lower performance than **NACNN+LogMels**. As for **NACNN+LogMels***, we got a better performance on RAVDESS than **NACNN+LogMels**.

G. Comparison with other methods

UAR is computed to compare with 3-D ACRNN [8] method on IEMOCAP and Emo-DB speech corpus. From Table V, Comparing with 3-D ACRNN in terms of UAR, **NACNN+LogMels** achieved 85.11% for Emo-DB, thus +2.3% improvement and for IEMOCAP, **NACNN+LogMels** result is very close to their method.

As shown in Table V, comparing **NACNN+LogMels** with GCNN [10] on IEMOCAP in terms of WAR, UAR and MAP, we computed WAR, UAR and MAP on IEMOCAP and achieved improvement of +5.97%, +4.76 and +4.27% respectively.

Comparing with baseline LMT [20], it achieved overall accuracy of 70% for seven emotions based on total test samples for RAVDESS corpus. **NACNN+LogMels** achieved overall accuracy of 75.25% for eight emotions on RAVDESS corpus based on 279 test samples total and recorded 5.25% as an improvement. For Emo-DB corpus, showed in Table V, LMT achieved accuracy 64.52% and **NACNN+LogMels** achieved 88.07%, thereby obtaining about +23.55% of improvement.

TABLE V
COMPARISON BETWEEN DIFFERENT METHODS ON EMO-DB, RAVDESS AND IEMOCAP

Methods		**Datasets**		
		Emo-DB	IEMOCAP	RAVDESS
3-D ACRNN [8]	UAR	82.82±4.99	**64.74±5.44**	N/A
GCNN [10]	UAR	N/A	60.59	N/A
	WAR	N/A	59.78	N/A
	WAP	N/A	61.48	N/A
LMT [20]	ACC	64.52	N/A	70
NACNN	UAR	82.79±4.39	62.40±0.73	71.98±3.53
	ACC	86.5	63.55	74.63
NACNN+LogMels	UAR	**85.11±3.80**	**64.35±0.54**	**73.64±1.87**
	ACC	**88.07**	**65.35**	**75.25**
	WAR	87.57	65.75	74.63
	WAP	88.57	65.75	74.75
NACNN+LogMels*	UAR	84.17±3.07	62.61±0.41	**74.49±1.72**
	ACC	87.24	63.72	**76.66**

V. CONCLUSION

This paper proposes a model based on 1D convolutional network for speech emotion recognition. Features were mainly extracted by complementary features, then based on 1D to capture emotional features. The model is evaluated on three databases, Emo-DB, RAVDESS and IEMOCAP and showed the superiority of the results in all, and more efficient than the baselines in SER. Experimental results demonstrate that 1D convolutional neutral networks can capture efficient emotional features for SER compared to other methods. And also proved that our method is efficient on SER and has a good performance in this area.

VI. ACKNOWLEDGEMENT

This work was supported by the Intelligence Engineering Lab (IELAB), University of Electronic Science and Technology of China.

References

[1] S. Basu, A. Bag, Mahadevappa M, J. Mukherjee, and R. Guha, "Affect detection in normal groups with the help of biological markers," in *2015 Annual IEEE India Conference (INDICON)*, pp. 1–6, Dec 2015.

[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, pp. 829–837, July 2000.

[3] S. Basu, J. Chakraborty, A. Bag, and M. Aftabuddin, "A review on emotion recognition using speech," in *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pp. 109–114, March 2017.

[4] P. Chandrasekar, S. Chapaneri, and D. Jayaswal, "Automatic speech emotion recognition: A survey," in *2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA)*, pp. 341–346, April 2014.

[5] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using cnn," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, (New York, NY, USA), pp. 801–804, ACM, 2014.

[6] Q. Kong, I. Sobieraj, W. Wang, and M. Plumbley, "Deep neural network baseline for dcase challenge 2016," in *Detection and Classification of Acoustic Scenes and Events 2016*, 2016.

[7] H. Phan, L. Hertel, M. Maass, P. Koch, R. Mazur, and A. Mertins, "Improved audio scene classification based on label-tree embeddings and convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 1278–1290, June 2017.

[8] 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, pp. 1440–1444, Oct 2018.

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

[10] J. Liu, W. Han, H. Ruan, X. Chen, D. Jiang, and H. Li, "Learning salient features for speech emotion recognition using cnn," in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pp. 1–5, May 2018.

[11] F. Burkhardt, A. Paeschke, M. A. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *INTERSPEECH*, 2005.

[12] C. Busso, M. Bulut, C. chun 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," 2008.

[13] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLOS ONE*, vol. 13, pp. 1–35, 05 2018.

[14] N. B. Wunarso and Y. E. Soelistio, "Towards indonesian speech-emotion automatic recognition (i-spear)," in *2017 4th International Conference on New Media Studies (CONMEDIA)*, pp. 98–101, Nov 2017.

[15] G. Liu, W. He, and B. Jin, "Feature fusion of speech emotion recognition based on deep learning," in *2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, pp. 193–197, Aug 2018.

[16] L. Guo, L. Wang, J. Dang, L. Zhang, and H. Guan, "A feature fusion method based on extreme learning machine for speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2666–2670, April 2018.

[17] S. Basu, J. Chakraborty, and M. Aftabuddin, "Emotion recognition from speech using convolutional neural network with recurrent neural network architecture," in *2017 2nd International Conference on Communication and Electronics Systems (ICCES)*, pp. 333–336, Oct 2017.

[18] B. Puterka and J. Kacur, "Time window analysis for automatic speech emotion recognition," in *2018 International Symposium ELMAR*, pp. 143–146, Sep. 2018.

[19] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1–4, Dec 2016.

[20] A. A. A. Zamil, S. Hasan, S. M. Jannatul Baki, J. M. Adam, and I. Zaman, "Emotion detection from speech signals using voting mechanism on classified frames," in *2019 International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST)*, pp. 281–285, Jan 2019.

[21] B. McFee, M. McVicar, S. Balke, V. Lostanlen, C. Thom, C. Raffel, D. Lee, K. Lee, O. Nieto, F. Zalkow, D. Ellis, E. Battenberg, R. Yamamoto, J. Moore, Z. Wei, R. Bittner, K. Choi, nullmightybofo, P. Friesch, F.-R. Stter, Thassilo, M. Vollrath, S. K. Golu, nehz, S. Waloschek, Seth, R. Naktinis, D. Repetto, C. F. Hawthorne, and C. Carr, "librosa/librosa: 0.6.3," Feb. 2019.

[22] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai, "Music type classification by spectral contrast feature," in *Proceedings. IEEE International Conference on Multimedia and Expo*, vol. 1, pp. 113–116 vol.1, Aug 2002.

[23] C. Harte, M. Sandler, and M. Gasser, "Detecting harmonic change in musical audio," in *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia*, AMCMM '06, (New York, NY, USA), pp. 21–26, ACM, 2006.

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015.