

# An Experimental Study of Speech Emotion Recognition Based on Deep Convolutional Neural Networks

W. Q. Zheng, J. S. Yu, Y. X. Zou\*

ADSPLAB/ELIP, School of Electronic Computer Engineering  
Peking University  
Shenzhen, China

\*correspondence author: zouyx@pkusz.edu.cn

**Abstract**—Speech emotion recognition (SER) is a challenging task since it is unclear what kind of features are able to reflect the characteristics of human emotion from speech. However, traditional feature extractions perform inconsistently for different emotion recognition tasks. Obviously, different spectrogram provides information reflecting difference emotion. This paper proposes a systematical approach to implement an effectively emotion recognition system based on deep convolution neural networks (DCNNs) using labeled training audio data. Specifically, the log-spectrogram is computed and the principle component analysis (PCA) technique is used to reduce the dimensionality and suppress the interferences. Then the PCA whitened spectrogram is split into non-overlapping segments. The DCNN is constructed to learn the representation of the emotion from the segments with labeled training speech data. Our preliminary experiments show the proposed emotion recognition system based on DCNNs (containing 2 convolution and 2 pooling layers) achieves about 40% classification accuracy. Moreover, it also outperforms the SVM based classification using the hand-crafted acoustic features.

**Keywords**—speech emotion recognition; deep convolutional neural networks; principle component analysis whitening; speech spectrogram

## I. INTRODUCTION

Despite the great progress made in artificial intelligence, we are still far from being able to naturally interact with machines, partly because machines do not understand our emotion states [1]. Emotions play an important role in the human-computer interaction. Recently, speech emotion recognition, which aims to analyze the emotion states through speech signals, has been attracting increasing attention. However, Speech emotion recognition is still a very challenging task for which how to extract effective emotional features is an open question [2][3].

Speech emotion recognition aims to identify the high-level effective emotion states of an utterance from the low-level features. It can be treated as a classification problem based on speech sequences. In order to achieve better emotion classification performance, many acoustic features have been investigated, such as energy-related features, pitch frequency features [4], formant frequency [5], Zero Crossing Rate (ZCR) [6], Linear Prediction Coefficients (LPC), Linear Prediction

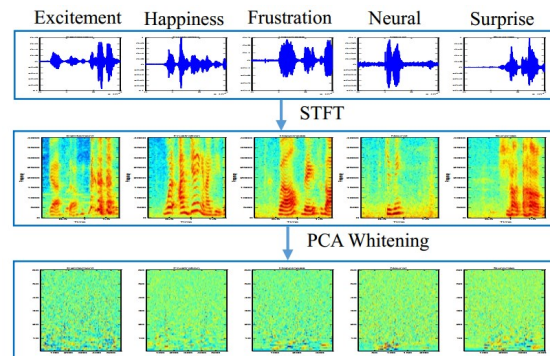


Fig.1. Example images: 1s chunks from speech data files corresponding to five emotion classifications. (Top channel represents speech signals, the middle one denotes the spectrograms, and the bottom one is the spectrograms after PCA whitening)

Cepstrum Coefficients (LPCC), Mel-Frequency Cepstrum Coefficients (MFCC) and its first derivative [7][8][9], and RASTA-PLP [10], etc. Some researchers also explored generative models, such as Gaussian mixture models (GMMs) and Hidden Markov models (HMMs), to learn the distribution of the extracted speech features, and then apply the Bayesian classifier or the maximum likelihood principle for emotion recognition [11][12]. Some trained the universal background models (UBMs) from the low-level speech features and then generated the supervectors for SVM classification [13][14], which is a widely used approach in speaker recognition. Besides, several classification methods, such as Neural Networks (NN) [15], K-nearest Neighbors (KNN) [16], decision trees [17], and Support vector machines (SVM) [18], have also been investigated in speech emotion recognition.

However, it is uncertain whether these hand-tuned feature sets can sufficiently and efficiently reflect the emotional characteristics of speech [2]. Moreover, their performance varies greatly in different databases and scenarios. Hence, the extraction of effective features reflecting emotions seems to be a difficult task. In many cases, the feature extraction has to be discussed based on different situation and chosen manually.

Fig.1 shows the example image from 1s chunks from speech data file corresponding to five emotion classifications. Specifically, it contains three channels. The top channel denotes speech signals, the middle one addresses the

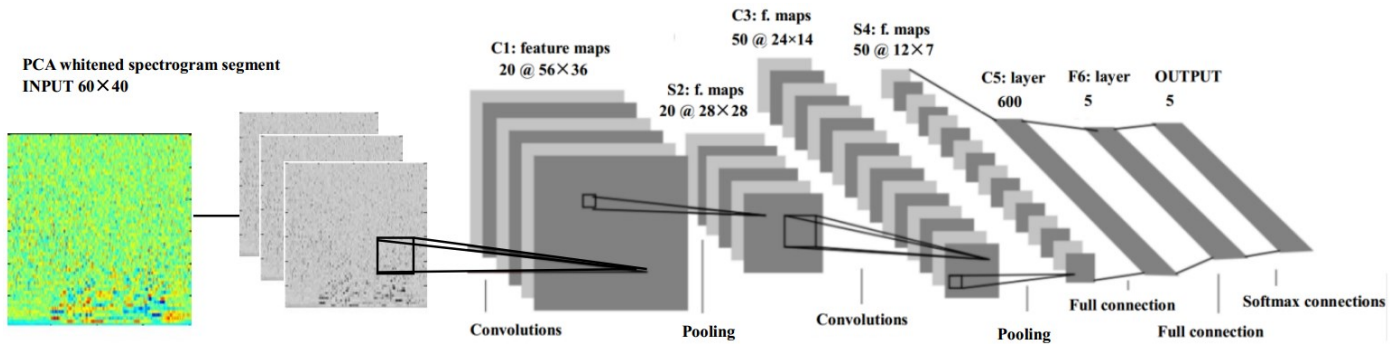


Fig.2. the DCNNs architecture for speech emotion recognition using the spectrogram segment as input

spectrograms, and the bottom one is the spectrograms after principal component analysis (PCA) whitening. It is noted that there are difference among spectrograms corresponding to five emotion classifications in color and texture. And the bottom one gives the difference in detail. Recent researches has applied the pattern recognition and machine learning based approach to solve the speech emotion recognition. Deep learning can be called ‘representation learning’ to distinguish the emotion-related features of speech spectrogram. Correspondingly, our study aims at applying the deep learning based method to solve the speech emotion recognition task. We try to train a deep learner using deep convolutional neural networks (DCNNs), which can learn the relevant, complex feature representation from short segments of speech data, to classify speech emotion. We hope to improve emotion classification through supervised training of a deep CNNs on speech spectrograms. In fact, our preliminary experimental results show the proposed DCNNs based speech emotion recognition approach outperforms the emotion classification on the hand-crafted acoustic features.

The remaining of this paper is organized as follows. In the next section, we relate our work to prior studies. And then we describe our proposed DCNNs based speech emotion recognition system in Section 3. the experimental results is presented and discussed in Section 4 and we conclude our study at last.

## II. RELATION TO PRIOR WORK

Deep learning is an emerging field in machine learning in recent years. A very promising characteristic of deep neural networks is that they can learn high-level invariant features from raw data [19][20], which is potentially helpful for emotion recognition. Recent studies show that CNNs have the capacity to learn higher order data features and have proven to be greatly successful models for classifying image data.

CNNs have been used specifically for the task of speech emotion recognition. Huang *et al.* [21] achieved good performance of speech emotion recognition by trying to learn salient feature maps using an auto-encoder followed by a CNN. However, no temporal model was utilized, because a linear classifier was trained across features from all time frames in a given speech file to predict the associated emotion state. We find this approach is not flexible as variable length input speech is not allowed. On the other hand, Namrata *et al.* [22]

presented a CNN model to extract features from speech data, and verified that emotion classification using features extracted from single-layer CNN model (1024 filters) outperformed classification on typically extracted acoustic features. But the spectrogram is log-transformed and subsampled the spectrograms to frequency range 1-2000 Hz, which may lose some important information for speech emotion classification.

Motivated by the special property of CNNs, we apply deep CNNs to learn the features for speech emotion recognition from audio spectrogram data, which can be viewed as an image representation of audio data along frequency and time axes. The spectrogram is further log-transformed and processed using PCA whitening (with 60 components) to reduce the dimensionality and some interference to the emotion classification task. As a result, our preliminary experiments verify the effectiveness of PCA whitening and show the proposed DCNNs based speech emotion recognition method outperforms the emotion classification using the hand-crafted acoustic features.

## III. THE PROPOSED SPEECH EMOTION RECOGNITION SYSTEM BASED ON DEEP CONVOLUTIONAL NEURAL NETWORKS

In this section, we firstly introduce the details of the architecture of the DCNNs. The input and output details of the each layer of DCNNs is addressed in the following first subsection. A block of the proposed speech emotion recognition system based on DCNNs is described at last.

### A. The DCNN Architecture

Following the CNN architecture proposed in [23], the DCNNs construction for speech emotion recognition (termed as DCNNs-SER in short) is illustrated in Fig.2, where there are two stages of convolution (conv.) and pooling (pool.) and two fully-connected (FC) layers.

**Convolutional layer:** PCA whitened spectrogram is obtained after the log-spectrogram is computed and PCA technique is conducted, which is viewed as a map to input into the DCNNs. Then we regard the input as a 3-D array with  $N$  2-D input maps size of  $n_1 \times n_2$ , the  $i^{th}$  input map is denoted as  $x_i$ . The output is also a 3-D array with  $M$  2-D output maps size of  $m_1 \times m_2$  and the  $j^{th}$  output map will be denoted as  $y_j$ . A filter (convolutional kernel)  $k_{ij}$  in the filter bank has the size of  $l_1 \times l_2$ ,

connects input map  $x_i$  to output map  $y_j$ . And the input-output model is based on the following equation:

$$y_j = b_j + \sum_i k_{ij} \otimes x_i \quad (1)$$

where  $b_j$  is a trainable bias parameter and  $\otimes$  is the 2-D convolution operator. Each filter detects a particular feature at every location on the input. Hence spatially translating the input of a feature detection layer will translate the output but leave it otherwise unchanged [24]. By sharing the equal values of the same filter, the number of coefficients to be trained is great reduced. Furthermore, the operation of local convolution is more like human vision receptive field. Thus, the feature extracted is more reasonable and effective.

**Pooling layer:** In this layer, the  $i^{th}$  output map  $y_i$  is corresponding to the  $i^{th}$  output map  $x_i$  of the previous layer. The input-output model is denoted as:

$$y_i = f(pool(x_i)) \quad (2)$$

where  $pool(.)$  computes the average value of the input map (average pooling) or the maximum value of the map (max pooling) over a small neighborhood ( $3 \times 3$  or  $5 \times 5$ ). The distance between successive receptive fields is larger than 1 (usually 3 or 5 according to the filter size). Hence, pooling operation will result in a low resolution map but with stronger robust to small variance in the corresponding location of the input map.  $f(.)$  is a *sigmoid* or *tanh* function that is applied point-wisely. Through the space sub-sampling operation of pooling, the feature will be more robust to small disturbance or noise in the map.

**Fully-connected layer:** This layer is similar to the traditional neural network, which connects each output unit to every input unit. The activation function is a *sigmoid* or *tanh* function. The input-output model is given as:

$$y_j = f\left(\sum_i W_{ij} x_i + b_j\right) \quad (3)$$

$$f(z) = 1/(1 + \exp(-z)) \text{ or } f(z) = \tanh(z) \quad (4)$$

where  $y_j$  and  $x_i$  are the output map and input map corresponding,  $W_{ij}$  and  $b_j$  are trainable weights and bias parameters. The final two fully-connected layers are actually a classifier, using the feature extracted by the layers before as input and accomplished the classification task.

In this design, the system shown in Fig.2 is trained in a supervised manner, where stochastic gradient descent method is employed to minimize the difference between the actual output and desired output at the fully-connected layer. In our study, the gradients are computed using the back-propagation (BP) algorithm. All the coefficients of all the filters and together with other parameters in every layer will be update during the training procedure [25].

### B. DCNNs Based Speech Emotion Recognition System

To illustrate the task of speech emotion recognition effectively, we describe the proposed DCNNs based speech emotion recognition system shown in Fig.3. The proposed algorithm is termed as PCA-DCNNs-SER in short, which is

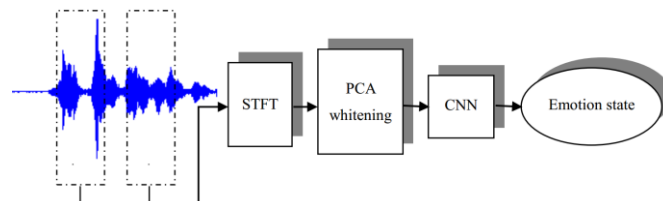


Fig. 3 Block diagram of the proposed DCNNs based speech emotion recognition system.

developed under the PCA whitening process of spectrogram from speech data and emotion prediction by DCNNs.

Specifically, we trained the DCNNs in a supervised learning manner by using a large speech dataset with the corresponding emotion labels. As shown in the Fig.3, we firstly extracted the spectrogram from each utterance of the speech dataset. The spectrogram has a 25 ms window size with 15 ms overlaps. Second, the spectrograms is further log-transformed and processed using PCA whitening (with 60 components), which is split into non-overlapping segments each representing 1ms of speech. Third, the PCA whitened spectrogram segments are fed into the DCNNs for learning the effective features. Finally, the speech emotion state can be successfully predicted through the DCNNs.

## IV. EXPERIMENTAL RESULTS AND ANALYSIS

### A. Experimental Setting

In order to evaluate our proposed approach, we use the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [26]. This database contains audiovisual data from 10 actors, and we only use audio track in our experiment. Each utterance in the database is labeled by three human annotators using categorical and dimensional labels. In our study We use categorical labels and only consider utterances with labels from five emotions: excitement, frustration, happiness, neutral and surprise. Since three annotators may give different labels for an utterance, we take those utterances which are given the same label by at least two annotators to avoid ambiguity in our experiment.

We train the model in the speaker-independent manner. we randomly choose 80% utterances of each emotion classification to construct the training dataset, and use the other 20% utterances for test. The length of an utterance is unequal and its mean is 4s. The detail of the dataset description is shown in Table I. The speech data sampled at 16 kHz is converted into frames with a 25-ms window sliding at 10-ms each time. A short-time Fourier transformation with 512 points is applied to

TABLE I. DATA DESCRIPTION OF IEMOCAP DATABASE

IEMOCAP Database			
Emotion classification	Total (utterances)	Training set (utterances)	Testing set (utterances)
Excitement	1038	830	208
Happiness	594	475	119
Frustration	1842	1473	369
Neutral	1707	1365	342
Surprise	107	85	22

compute the DFT of each overlapping Hamming windowed frame. The size of the spectrogram segment level feature is set to 40 frames. So the total length of a segment is  $10\text{ms} \times 40 + (25-10)\text{ms} = 385\text{ms}$ . It is clear that emotional information is usually encoded in one or more speech segments whose length varies on many factors such as speakers and emotions. The appropriate segment length for emotion recognition is still an open problem. Luckily, It is told that a speech segment longer than 250ms contains sufficient emotional information [16][27]. Then, the spectrogram segment is transformed into log-power spectra feature with  $256 \times 40$  dimensions, which is further processed through Principal Component Analysis (PCA) whitening (with 60 components) [21]. So the input of DCNNs is data matrix ( $60 \times 40$  dimensions) which is normalized into  $[-1, 1]$ . In our experiment, the DCNNs contains two stages of convolution and pooling, and two fully-connected layers. Finally, a softmax classifier is applied to predict the emotion classification. The details of the parameters setting of DCNNs is shown in Fig.2.

### B. Results And Discussion

In order to validate the performance of the proposed PCA-DCNNs-SER, we consider the standard hand-crafted acoustic features for speech emotion recognition task. We have extracted 85 dimensional hand-crafted features from the speech files for each 25ms frame with 10ms sliding. which have been used in past studies to identify the speech emotion. The features are described briefly in Table II. We take SVM to classify the emotion from the input speech data based on 85 dimensional hand-crafted features described in Table II. After the coarse hyperparameter tuning, the best SVM model achieves 37.61% classification accuracy on the test set. From the results in the Table III, it is encouraged to find the best classification accuracy is 40.02% by the proposed PCA-DCNNs-SER with the CNN architecture, containing 2 convolution and 2 pooling layers. Obviously, the speech emotion classification using CNNs with the input data of PCA whitened spectrogram segments can obtain better performance than the SVM classification based hand-craft features. Besides, the PCA whitening process of spectrogram proves to be effective and boost the classification accuracy of speech emotion. The DCNNs based SER without PCA whitening (termed as DCNNs-SER) acts worse than PCA-DCNNs-SER. Moreover, we also conduct the experiments to analyze the effect of the architecture of DCNNs. We find the performance of DCNNs-SER or PCA-DCNNs-SER with different layers obtain different classification accuracy. It is shown in Table III that the PCA-DCNNs-SER (containing 2 convolution and 2 pooling layers) achieves the best accuracy among others.

Generally speaking, the emotion classification accuracy is unable to achieve better on this database, since the data distribution of each emotion from the database is imbalanced (see Table I). Just as the results shown in Table III, the imbalanced databased leads to the difficulty for more accurate classification. It also limits the performance improvement of the DCNNs architecture, though increasing the number of the convolution or pooling layers. Hence, it is clear to predict a better speech emotion classification accuracy can be achieved on a balanced database.

TABLE II. ILLUSTRATION HAND-CRAFTED ACOUSTIC FEATURES

Feature	Num.	Description
MFCC+ $\Delta$ + $\Delta\Delta$	39	Timbre characterized by 39 coeff.
LPC	16	Speech parameters for acoustic model
LPCC	16	Derived from LPC
RASTA-PLP	12	Robust to linear spectral distortions
Pitch	1	Fundamental freq. of sound
Zero crossing	1	Noise in sample
<b>whole</b>	<b>85</b>	

TABLE III. SPEECH EMOTION CLASSIFICATION ACCURACY FOR SVM MEDEL AND DIFFERENT CNNs ARCHITECTURES

Approach	Test acc
hand-crafted acoustic features + SVM	0.3761
Log Spec+CNN (1 Conv.+1 pool.)	0.3598
Log Spec+ CNN (2 Conv.+2 pool.)	0.3371
Log Spec+CNN (3 Conv.+2 pool.)	0.3351
Log Spec+PCA whiten+CNN (1 Conv.+pool.)	0.3524
Log Spec+PCA whiten +CNN (2 Conv.+2 pool.)	<b>0.4002</b>
Log Spec+PCA whiten +CNN (3 Conv.+2 pool.)	0.3807

### V. CONCLUSION

In this paper, a deep convolution neural networks based approach has been developed to learn the effective features for speech emotion recognition from audio spectrogram data, which is further log-transformed and processed using principle component analysis (PCA) whitening. Accordingly, a speech emotion recognition algorithm termed as PCA-DCNNs-SER is proposed. Preliminary experiments have been conducted to evaluate the performance of PCA-DCNNs-SER on the IEMOCAP database. Results show that our proposed PCA-DCNNs-SER (containing 2 convolution and 2 pooling layers) is able to obtain about 40% classification accuracy, which outperforms the SVM based SER using hand-crafted features. Besides, the PCA whitening process of spectrogram proves to be effective and boost the performance of DCNNs based SER. Our future work is to combine feature extraction and context learning for better emotion prediction from audio data of variable length with a large data.

### ACKNOWLEDGMENT

This work is partially supported by National Natural Science Foundation of China (No: 61271309) and Shenzhen Science & Technology Research Program (No: JC201105170727A).

### REFERENCES

- [1] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," *Proceedings of INTERSPEECH, ISCA, Singapore*, pp. 223-227, 2014.
- [2] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, pp. 572-587, 2011.
- [3] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, pp. 1062-1087, 2011.
- [4] D. Ververidis, C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification," in *Acoustics, Speech, and Signal Processing*,

2004. *Proceedings (ICASSP'04). IEEE International Conference on*, 2004, pp. 1-593-6 vol. 1.
- [5] Z. Xiao, E. Dellandrea, W. Dou, and L. Chen, "Features extraction and selection for emotional speech classification," in *Advanced Video and Signal Based Surveillance, 2005. AVSS 2005. IEEE Conference on*, 2005, pp. 411-416.
- [6] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell System Technical Journal*, vol. 54, pp. 297-315, 1975.
- [7] T.-L. Pao, Y.-T. Chen, J.-H. Yeh, and P.-J. Li, "Mandarin emotional speech recognition based on SVM and NN," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, 2006, pp. 1096-1100.
- [8] Y. Pan, P. Shen, and L. Shen, "Feature Extraction and selection in speech emotion recognition," in *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 2005), Como, Italy, 2005*.
- [9] P. Shen, Z. Changjun, and X. Chen, "Automatic speech emotion recognition using support vector machine," in *Electronic and Mechanical Engineering and Information Technology (EMEIT), 2011 International Conference on*, 2011, pp. 621-625.
- [10] H. Morgan, N. Bayya, A. Kohn, and P. Hermansky, "RASTA-PLP speech analysis," ICSI Technical Report TR-911991.
- [11] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *Proceedings of IEEE ICASSP 2003*, vol. 2. IEEE, 2003, pp. II-1.
- [12] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, *et al.*, "Emotion recognition based on phoneme classes," in *Interspeech, 2004*, pp. 205-211.
- [13] H. Hu, M.-X. Xu, and W. Wu, "GMM Supervector Based SVM with Spectral Features for Speech Emotion Recognition," in *ICASSP (4), 2007*, pp. 413-416.
- [14] T. L. Nwe, N. T. Hieu, and D. K. Limbu, "Bhattacharyya distance based emotional dissimilarity measure for emotion classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7512-7516.
- [15] X. Mao, L. Chen, and L. Fu, "Multi-level speech emotion recognition based on HMM and ANN," in *Computer Science and Information Engineering, 2009 WRI World Congress on*, 2009, pp. 225-229.
- [16] Y. Kim and E. M. Provost, "Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 3677-3681.
- [17] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, pp. 1162-1171, 2011.
- [18] Y. Chavhan, M. Dhore, and P. Yesaware, "Speech emotion recognition using support vector machine," *International Journal of Computer Applications*, vol. 1, pp. 6-9, 2010.
- [19] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, pp. 1798-1828, 2013.
- [20] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks-studies on speech recognition tasks," *arXiv preprint arXiv:1301.3605*, 2013.
- [21] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in *Proceedings of the ACM International Conference on Multimedia*, 2014, pp. 801-804.
- [22] N. Anand and P. Verma, "Convolved Feelings Convolutional and recurrent nets for detecting emotion from audio data."
- [23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.
- [24] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, 2010, pp. 253-256.
- [25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278-2324, 1998.
- [26] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335-359, 2008.
- [27] E. M. Provost, "Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 3682-3686.