End-to-End Speech Emotion Recognition Using Multi-Scale Convolution Networks

Tatinati Sivanagaraja*, Mun Kit Ho*, Yubo Wang†, and Andy W H Khong*
† School of Life Science and Technology, Xidian University, Xi'an, China.
* School of Electrical Engineering, Nanyang Technological University,
Singapore (*e-mail : *tatinati@ntu.edu.sg*)

***Abstract*—Automatic speech emotion recognition is one of the challenging tasks in machine learning community mainly due to the significant variations across individuals while expressing the same emotion cue. The success of emotion recognition with machine learning techniques primarily depends on the feature set chosen to learn. The formulation of appropriate feature set that cater all the variations in emotion cues however is not a trivial task. Recent works on emotion recognition with deep learning techniques thus focus on the $end-to-end$ learning scheme which identifies the features directly from the raw speech signal instead of relying on hand-designed feature set. The existing methods in this scheme however did not take into account the fact that speech signals often exhibit significant features at different time scales and frequencies than in the raw form. To address this issue, in this work, an $end-to-end$ neural network model named as Multi-scale Convolution Neural Network (MCNN) is proposed to automatically identify the features at different time scales and frequencies of the raw speech signal. The proposed model further leverages on the multi-branch input layer and tunable convolution layers to learn the identified features and then recognizes the emotion cues accordingly. The MCNN method is evaluated with the SAVEE emotion database and results highlight that the proposed method improves the emotion recognition accuracy significantly as compared to the existing methods.**

I. Introduction

Deep neural networks achieved great results across many automatic speech recognition applications from audio-visual, speaker and speech recognition, and its emotion [1]. Numerous studies have shown that a series of neural network architectures such as autoencoders [2], convolution neural networks (CNN) [3], long short-term memory (LSTM) [4] models have enough degrees of freedom to model inherent non-stationary nature of the speech signals. Despite having the deep structure, applications related to speech domain generally employ hand-engineered features (like Mel-frequency coefficients) as input features to achieve the task at hand. Identifying appropriate features and then optimizing the network based on the features for every application however is demanding and non-trivial task. The recent emerging trend in machine learning community therefore is deriving the representations from the raw input signal directly [1], [5]. The motive is that the deep network ultimately identifies and learns the appropriate features that give better performance from the raw signal directly for every task.

Speech Emotion Recognition (SER) is one of the challenging tasks for machine learning community over the decades because the same emotion can be expressed in numerous ways by different people. As such, the generic features that can distinguish the emotions are unclear. To overcome this, researchers have developed a plethora of methods for emotion recognition from speech signals with hand-designed features. CNN based methods for SER has been developed in [3] to learn the salient features for SER. This method however selected optimal features from a pool of hand-designed features. Later, LSTM based methods with spectrogram as input feature were developed for SER in [4]. The end-to-end approach (raw speech signals as input features) based emotion recognition is recently adapted for SER in [5]. These methods however struggle with memory and training instability (due to the perturbations in raw speech signals) concerns, because its not always optimal to use very large sequences as input feature for training the network. Further, speech signals generally contain emotion-related features in different time scales and frequency components rather than in raw form. To address these issues, in this work, a multi-scale convolution network (MCNN) based end-to-end model is proposed for SER applications.

The proposed MCNN method initially performs three transformations on the raw speech signal to decompose the signal into multiple temporal and spectrum scales. Tunable local and global convolution layers are employed subsequently to identify and learn the features from each decomposed time series. With the multi-branch input layer, the convolution layers and followed by the fully connected network, the proposed MCNN method is capable of learning the temporal patterns in speech signal as well as the features that exists in the multiple frequency components. The proposed method is evaluated with SAVEE emotion database and results showed the MCNN improves the emotion recognition performance compared to the existing methods.

The paper is organized as follows. In Section II, the background theory for multi-scale convolution networks for speech emotion recognition is provided. Section III presents a description about SAVEE emotion database and the performance evaluation of the proposed methods with this dataset. Section IV concludes the paper.

II. Methods

The main objective of this study is to investigate the suitability of multi-scale convolution networks for identifying the emotion cues that exists at different scales and frequencies in speech signals. In what follows, is a detail description

about the multi-scale convolution networks and its approach for speech emotion recognition.

Recognition of emotion cues from the raw speech signals ($\mathbf{s} = \{s_1, s_2, \cdots, s_n\}$, where n represents the number of samples) can be considered as a classical learning problem of estimating an unknown relationship between the elements of an input feature space ($\mathbf{S} \in \mathcal{R}^n$) and elements of a target space ($\mathbf{T} \in \mathcal{Z}$). The elements in the input feature space are formulated by considering the raw speech signal that belong to different emotion cues, can be given as $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_N]$, where N is the number of raw speech signals considered for training. The elements in target space are the class labels to which the corresponding elements input feature space $\mathbf{S}$ belong to, can be given as $\mathbf{T} = \{0, 1, \cdots, c\}$, where the c is the number of emotion cues. The formulated input vector and target vector with the training data are provided to the multi-scale convolution network to automatically identify and then learn the features at different scales that better represents the relationship between the input feature space and the target vector space, as shown in Fig.1. The non-linear map obtained after the training phase is stored and will be used to predict the emotion cues for raw speech signals in the testing phase.

A. Multi-scale Convolution Network (MCNN)

The architecture of multi-scale convolution network employed in this work for emotion recognition is depicted in Fig. 1. The MCNN framework mainly has three stages: transformation stage, local convolution stage, and global convolution stage. In the following subsections, a brief description about each stage is provided, for more details please refer to [6].

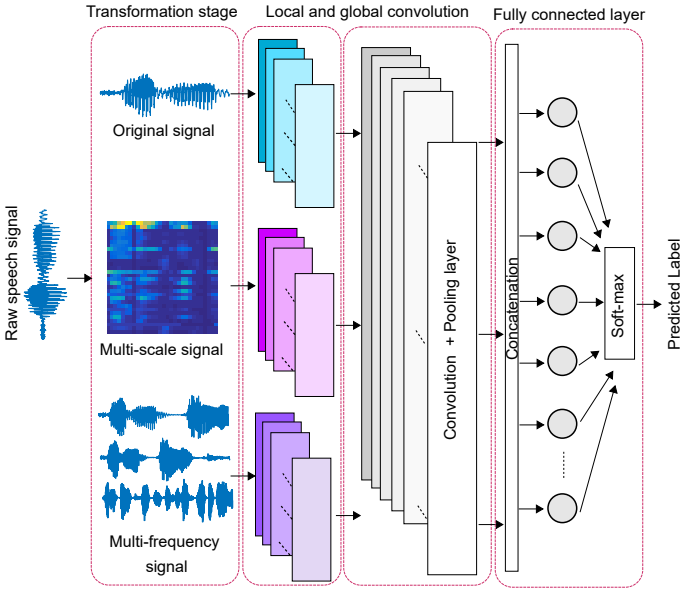

Fig. 1. Architecture of Multi-scale convolution network

1) Transformation Stage: Raw speech signals generally exhibit repeating deterministic pattern as well as random perturbations. The stochastic component consists of random variations in timing and amplitude of the signal. Small perturbations are called jitter and shimmer, but there are as well larger perturbations or trends that can be observed during the utterance of sustained vowels, called flutter. These random variations in frequency and amplitude hinders the emotion recognition efficacy from raw signal. To decompose the speech signal according to the variations in amplitude and frequency the following transformations are chosen in this work: 1) identity mapping, 2) multi-scale decomposition, and 3) multi-frequency decomposition.

Multi-scale decomposition: This transformation captures the temporal patterns that exists in the raw speech signal by decomposing the signal into multiple scales. Generally, the long-term temporal patterns exhibit over-all trend whereas the short-term temporal patterns exhibit subtle variations in local regions. Both these patterns are vital for proper identification of emotion from speech signals. In this work, we employed down-sampling to generate the multiple scales of the raw speech signal. The down-sampled signal at rate of p can be as given as

$$s_{ds} = \{s(1 + p * i)\}, i = 0, \cdots, \lfloor \frac{n-1}{p} \rfloor \quad (1)$$

where s_{ds} is the down sampled signal, s is the actual signal, and n is the number of samples. The whole raw speech signal and the segments (of length 3000 samples) down sampled signals at different rates is shown in Fig. 2 for illustration. The multi-scale speech signals obtained with this transformation contains information of the long-term and short-term temporal patterns, as depicted in Fig. 2(a)-(d), yield to better feature extraction as compared to raw speech signal alone.

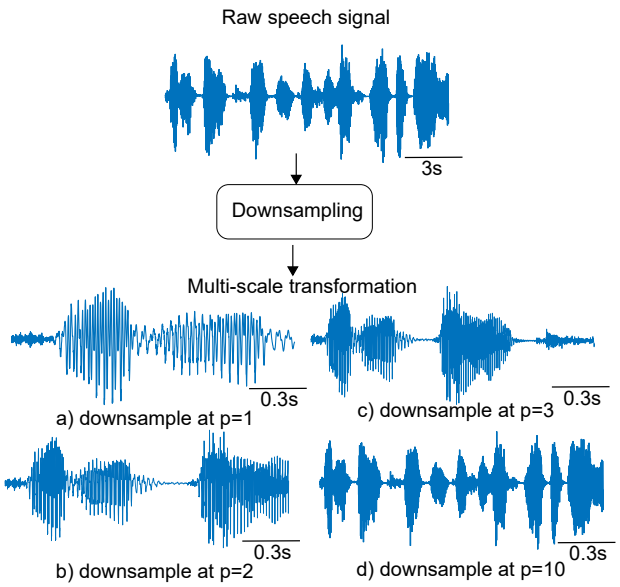

Fig. 2. Architecture of Multi-scale convolution network

Multi-frequency decomposition: This transformation decomposes the raw speech signals into its constituent frequency components. Further, this transformation will help identifying

features in the high-frequency perturbations. In this work, moving average filter with multiple degrees of smoothness is employed to decompose the raw speech signal. These decomposed speech signals will be learnt with the local convolution layer to identify the features. The moving average filter generates a new speech signal (s_{ma}) from raw speech signal (s) by the following equation:

$$s_{ma} = \frac{s_i + s_{i+1} + \cdots + s_{i+l-1}}{l} \tag{2}$$

where l is window size and $i = 0, 1, \cdots, n - l + 1$.

2) Local Convolution Stage: The speech signals obtained after the transformation stage is provided to 1-D local convolution filters followed by max pooling to extract the local and independent features from each generated time series.

The key component of local convolution is 1-D convolutions that operate on the transformed speech signals, can be given as:

$$(f * s)(t) = \sum_{k=-T}^{T} f(t)s(t-k) \tag{3}$$

where $f(\cdot)$ is the filter weights that are learnt from the emotion cues. The filter size chosen for local convolution in MCNN is the same for all the transformed speech signals. By selecting the same filter size, due to the down sampling and other employed transformations each local convolution stage captures features from different receptive fields. Max pooling with multiple sizes is performed on the outputs obtained from the local convolutions.

3) Global Convolution Stage: The features identified from each transformed time series are concatenated and provided to a convolution layer, named as global convolution stage. Similar to local convolution stage, 1-D convolutions are performed followed by the max pooling. The features obtained after the global convolution stage are provided to a fully-connected layer followed by soft-max to predict the emotion labels. In this work, we adopted the deep concatenation technique in [6] to concatenate all the feature maps vertically.

The MCNN employs cross-entropy loss function to train the network, can be given as:

$$max \sum_{i=1}^{N} log T_{s_i}^{(i)} \tag{4}$$

where $T_{s_i}^{(i)}$ is the s_i predicted label of instance i through MCNN. The parameters in local, global convolution layers, and fully connected layers ($\mathbf{W}$ and $\mathbf{b}$) are learnt through back propagation.

III. Results and Discussions

Firstly, a brief description about the SAVEE emotion database and the implementation procedure for evaluation of MCNN are provided. Later, the performed comparison analysis and the obtained inferences are detailed.

A. Speech Database

Surrey Audio-Visual Expressed Emotion (SAVEE) database is employed in this work to evaluated the emotion recognition with the MCNN method. The database consists of recordings from four male actors expressed in seven different emotions. In this work, as a proof-of-concept, we chose four emotions data (happiness, anger, fear, sadness) per subject to perform the classification. On a whole, among four subjects, there are 240 utterances of text in total. All the recordings are acquired at 44.1kHz with approximately 3 seconds long. For more details about the data acquisition and the preferences of uttered text, refer to [7]

B. Implementation

The raw speech signals at 44.1kHz are re-sampled to 11.05kHz and then each recording is segmented to reduce the dimension of the input feature vector for MCNN. The segmentation is performed in two ways: a) segment of length (L) without any overlapping and b) segment of length (L) with overlapping of $L/2$. Since the dimension of data required to accurately identify the emotion from speech is still an open research topic, in this work, we chose various lengths of segments for analysis. The chosen lengths are 1100 samples (100 ms), 2500 samples (250 ms), 3500 samples (350 ms), 5000 (500 ms), and 7000 samples (700 ms). Each segment is given the same label as their source speech signal. While training MCNN, all the segments are considered as independent training inputs. This segmentation further reduced memory and training instability concerns.

Hyper parameters selection: The optimal performance of MCNN requires optimal initialization of hyper parameters. The list of MCNN hyper parameters and the initialization used in this work is tabulated in Table. I.

TABLE I
THE MCNN HYPER PARAMETERS AND THEIR INITIALIZATION

Parameter	Initialization values
Convolution filter size	5×1
Pool Factor	2×1
Activation function	Tanh
Optimizer	Stochastic gradient descent
Learning rate	0.1

Further, in this work, the simulations are carried out for 200 epochs with training batch size of 10.

C. Comparison Analysis

The dimension of input feature vector is one of the major concerns for emotion recognition with end-to-end learning approach. The longer sequences raise memory and training instability concerns whereas with smaller sequences might not contain all the information to identify the emotion. In order to choose the best dimension for input feature vector, the raw speech signal was segmented with various lengths (as mentioned in Section III-B). With the employed segmentation, in this work, we formulated five data sets and experiments are conducted on each data set separately. In each data set 56%

of data is used as training data set, 14% as validation data set and the remaining 30% of data as testing data set.

TABLE II
EMOTION RECOGNITION ACCURACY WITH VARIOUS NON-OVERLAPPING SEGMENT LENGTHS

Segment length	Validation accuracy	Testing Accuracy
1100 samples (100 ms)	46.64%	41.9%
2500 samples (250 ms)	46.74%	43.75
3500 samples (350 ms)	50%	45 %
5000 samples (500 ms)	43.58%	39.8%
7000 samples (700 ms)	43.82%	46.08%

Validation accuracy and testing accuracy obtained for both non-overlapping and 50% over-lapping segments are tabulated in Table. II & III respectively. The results mainly highlight that the input features formulated with overlapping segments yield better classification accuracy as compared to its counterpart. The major reason for this trend is, with non-overlapping segments there are so many discrepancies in input feature vector for the same label whereas with overlapping segments there discrepancies got reduced and hence the prediction accuracy improved. Furthermore, it can be inferred from the results that larger input feature vector improves the prediction accuracy, but with the small pool of training dataset may lead to over fitting.

TABLE III
EMOTION RECOGNITION ACCURACY WITH VARIOUS 50% OVERLAPPING SEGMENT LENGTHS

Segment length	Validation accuracy	Testing Accuracy
1100 samples (100 ms)	41..01%	41.36%
2500 samples (250 ms)	47.107%	47.21%
3500 samples (350 ms)	52.04%	49.5%
5000 samples (500 ms)	48.39%	47.74%
7000 samples (700 ms)	47.95%	50.28%

Compared to the existing methods in the literature [3] which used raw speech signals from SAVEE data sets as features for identifying the emotion cues, the proposed MCNN method improved the prediction accuracy by nearly 8%. This improvement underpins that the employed transformation stage represented the input feature vector in an elegant way for local and global convolution stages to identify and learn the features accurately as compared to the conventional learning from raw speech signals. Although, MCNN improves the prediction accuracy, identifying generic dimension for input feature vector needs further research.

IV. CONCLUSIONS

In this paper, we propose a multi-scale convolution based end-to-end speech emotion recognition method. To represent the perturbations in raw speech signals in an elegant way, transformation stage is introduced before the convolution stage. Our experiments on the SAVEE emotion dataset show that our model achieves significantly better performance in the test set in comparison to other existing models, thus demonstrating the efficacy of learning features through the multi-scale networks.

ACKNOWLEDGMENT

REFERENCES

[1] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," *Neural Networks*, 2017.

[2] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, 2014.

[3] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[4] W. Lim, D. Jang, and T. Lee, "Speech Emotion Recognition using Convolutional and Recurrent Neural Networks," *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1–4, 2016.

[5] G. Trigeorgis, F. Ringeval, R. Brueckner, and E. Marchi, "Adieu features ? end-to-end speech emotion recognition using a deep convolution recurrent network," in *International Conference on Acoustics, Speech, and Signal Processing*, 2016, pp. 5200–5204.

[6] Z. Cui, W. Chen, and Y. Chen, "Multi-scale convolutional neural networks for time series classificatio," *arXiv*, 2016.

[7] S. Haq and P. J. B. Jackson, "Speaker-dependent audio-visual emotion recognition," in *In Proc. Int'l Conf. on Auditory-Visual Speech Processing*, 2009, pp. 53–58.