INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Pyramid Memory Block and Timestep Attention for Speech Emotion Recognition

Miao Cao[1], *Chun Yang*[1], *Fang Zhou*[1,*], *Xu-cheng Yin*[1,*]

[1]Department of Computer Science and Technology, University of Science and Technology Beijing, Beijing, China

miaocao@xs.ustb.edu.cn, chunyang@ustb.edu.cn, zhoufang@ies.ustb.edu.cn, xuchengyin@ustb.edu.cn

Abstract

As a sequence model, Deep Feedforward Sequential Memory Network (DFSMN) has shown superior performance on many tasks, such as language modeling and speech recognition. Based on this work, we propose an improved speech emotion recognition (SER) end-to-end system. Our model comprises both CNN layers and pyramid FSMN layers, where CNN layers are added at the front of the network to extract more sophisticated features. A timestep attention mechanism is also integrated into our SER system, which makes the system learn how to focus on the more robust or informative segments in the input signal. Furthermore, different from traditional SER systems, the proposed model is applied directly to spectrograms which contain more raw speech information, rather than well-established hand-crafted speech features such as spectral, cepstral and pitch. Finally, we evaluate our system on the Interactive Emotional Motion Capture (IEMOCAP) database. The experimental results show that our system achieves 2.67% improvement compared to the commonly used CNN-biLSTM model which requires much more computing resource.

Index Terms: Speech emotion recognition, DFSMN, attention mechanism

1. Introduction

Thanks to recent progress of speech recognition technology and accompanying wide availability of speech recognition devices, human-machine speech interaction is being more common in our daily life. Such devices, however, can only recognize the word-level content of human speech but not emotions, while automatic emotion recognition has many useful applications for humancentric services and human-machine interactions, such as intelligent service robotics, automated call centers, and remote education.

A typical Speech Emotion Recognition (SER) system takes a speech waveform as input then outputs one of the target emotion categories. Traditional SER systems that utilize Gaussian Mixture Models (GMMs) [1, 2, 3], Hidden Markov Models (HMMs) [4, 5], Support Vector Machines (SVMs) [6, 7, 8] and Long Short-Term Memory (LSTM) [9, 10, 11] rely on well-established hand-crafted speech features. These features usually include spectral, cepstral, pitch, and energy features of the speech signal at the frame level [12]. Statistical functions of these features are then applied across multiple frames to obtain an utterance-level feature vector.

With the explosive development of deep learning technology, some researchers explored deep learning methods to build robust models for the SER task. In [13], an auto-encoder with Long Short-Term Memory (LSTM) neural networks via feature enhancement is proposed towards robust emotion recognition from spontaneous speech. Correspondingly, the Recurrent Neural Network (RNN) has been proved to have strong sequential modeling ability, especially for the speech recognition task. However, the training of RNN relies on backpropagation through time (BPTT) [14], which may bring problems such as more time consuming, gradient vanishing and exploding [15] due to its complex computation. To solve these problems, a Feedforward Sequential Memory Network (FSMN) is proposed [16]. Recently, lots of researches have proved that FSMN could model long-term relationship without any recurrent feedback [17] in many tasks, such as speech recognition and language model. Moreover, to build a very deep neural architecture, skip connection is applied to FSMN [18], which makes the large improvement to previous models.

Research activities in SER can be traced back to the 1980s [19, 20]. But SER is still challenging in real applications due to a range of factors such as gender, speaker, language and recording environment variations [21]. Many researchers try to solve these problems by designing well-established hand-crafted speech features to strengthen their connection with human emotions. However, these hand-crafted speech features are only useful for specific tasks, not universal. Based on the intuition that it is beneficial to emphasize the expressive part of the speech signal for emotion recognition. Especially, speech contains a lot of useful and irrelevant information at the same time. We propose an timestep attention mechanism to automatically select the useful information which reduces the impact of human factors in feature selection.

In this paper, we propose a DFSMN based deep learning approach for speech emotion recognition. A pyramid structure is applied in memory blocks. This improvement makes top layers contain more context information than bottom layers, which not only employs appropriate time dependency but also reduces the number of parameters. Inspired by computer vision tasks, we also deploy several CNN layers as the front-end to extract more sophisticated features from the original speech spectrograms. Besides, we introduce attention mechanism before the output to make the model emphasize expressive part of the speech signal for emotion recognition, which lets the model more robust.

The rest of the paper is organized as follows. FSMN and DFSMN are described in section 2. Section 3 presents our approach. In Section 4, we describe the experiment settings and present the evaluation results. Finally, we draw conclusions in Section 5.

[*] Corresponding author

http://dx.doi.org/10.21437/Interspeech.2019-3140

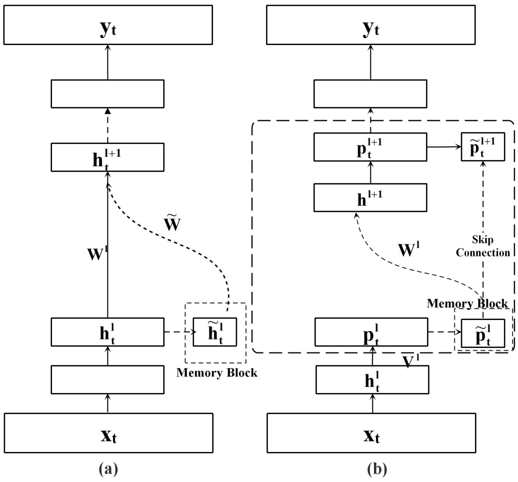

Figure 1: *FSMN(a) and DFSMN(b) architecture.*

2. FSMN

As shown in Figure 1 (a), FSMN [16] is a standard feedforward fully connected neural network appended with memory blocks in hidden layers. We can use a tapped-delay structure to encode $\boldsymbol{h_t}$, its N_1 previous time steps and N_2 next time steps into a fixed-sized representation, then computes their block-sum as current output:

$$\tilde{\boldsymbol{h}}_{\boldsymbol{t}} = f(\sum_{i=0}^{N_1} a_i \cdot \boldsymbol{h_{t-i}}) + f(\sum_{j=1}^{N_2} b_j \cdot \boldsymbol{h_{t+j}}) \tag{1}$$

where a_i from an N_1-dimension learnable vector $\boldsymbol{a} = \{a_0, a_1, \cdots, a_{N_1}\}$, b_j from an N_2-dimension learnable vector $\boldsymbol{b} = \{b_1, b_2, \cdots, b_{N_2}\}$, and $f(\cdot)$ is the activation function (sigmoid or ReLU). Furthermore, $\tilde{\boldsymbol{h}}_{\boldsymbol{t}}$ may be fed into next hidden layer in the same way as $\boldsymbol{h_t}$.

As shown in Figure 1(b), different from the original FSMN architecture, DFSMN [18] removes the direct forward connection between hidden layers, and take memory block as the only input. Meanwhile. skip connection and the memory strides are introduced to overcome the gradient vanishing and exploding problems.

The DFSMN component can be described as the following formulations:

$$\begin{aligned} \tilde{\boldsymbol{p}}_{\boldsymbol{t}}^{\boldsymbol{l}} = \mathcal{H}(\tilde{\boldsymbol{p}}_{\boldsymbol{t}}^{\boldsymbol{l-1}}) + \boldsymbol{p_t^l} + \sum_{i=0}^{N_1^l} \boldsymbol{a_i^l} \odot \boldsymbol{p_{t-s_1*i}^l} \\ + \sum_{j=1}^{N_2^l} \boldsymbol{b_j^l} \odot \boldsymbol{p_{t+s_2*j}^l} \end{aligned} \tag{2}$$

$$\boldsymbol{h_t^{l+1}} = f(\boldsymbol{W^l} \tilde{\boldsymbol{p}}_{\boldsymbol{t}}^{\boldsymbol{l}} + \boldsymbol{b^{l+1}}) \tag{3}$$

Here, $\boldsymbol{p_t^l} = \boldsymbol{V^l h_t^l} + \boldsymbol{b^l}$ denotes the linear output of the l-th linear hidden layer with the activation function (sigmoid or ReLU). $\tilde{\boldsymbol{p}}_{\boldsymbol{t}}^{\boldsymbol{l}}$ denotes the output of the l-th memory block. $\mathcal{H}(\cdot)$ denotes the skip connection within the memory block, which can be any linear of nonlinear transformation. In our work, we just use the identify mapping as following for all experiments:

$$\tilde{\boldsymbol{p}}_{\boldsymbol{t}}^{\boldsymbol{l-1}} = \mathcal{H}(\tilde{\boldsymbol{p}}_{\boldsymbol{t}}^{\boldsymbol{l-1}}) \tag{4}$$

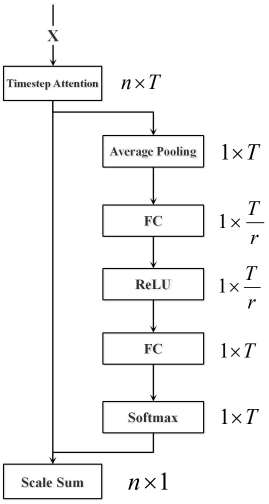

Figure 2: *Timestep attention mechanism architecture, in which n denotes the dimension of feature for each timestep, T denotes the total length of timestep, and r denotes the scale of the number of hidden layer nodes.*

N_1^l denotes the look-back order of the l-th memory block while N_2^l denotes look-ahead order. s_1 and s_2 denote strides of look-back order and strides of look-ahead order respectively. $\boldsymbol{a_i^l}$ and $\boldsymbol{b_i^l}$ are the coefficient vectors in memory blocks. But we just use scalar coefficients in our work as sFSMN [18].

3. Our Approach

As shown in Figure 3, our model contains three main components: CNN layers, pyramid FSMN layers and attention mechanism. Then, we'll explain them separately.

3.1. Pyramid Memory Block

In [16, 18], the memory block lengths are identical, which means you take the same value of N_1^l, N_2^l, s_1 and s_2 in Equation 2 through all hidden layers. With the memory block structure, the bottom layers not only extract long context information at certain time step t but also contain this information. Thus, the long-term relationship is duplicated and no longer needed in top layers [16] while top layers could get these relationship from bottom layers directly. In our approach, we proposed a pyramidal memory block structure, in which the model extracts more context information in deeper layers. This structure is implemented by increasing N_1^l, N_2^l, s_1 and s_2 with the layers go deeper. Hence, the bottom layers extract features from subtle information such as speed and rhythm of speech, while top layers extract features from more abstract information such as emotion and gender. This structure improves accuracy and reduces the number of parameters simultaneously.

3.2. Timestep Attention Mechanism

Attention mechanism in neural networks is inspired by the biological visual attention mechanism found in nature. For example, human beings are able to focus on a certain region of an image with high resolution while nearly ignoring the surrounding regions [10]. Furthermore, the region of focus can be shifted dynamically in a seemingly effortless manner. And we can get

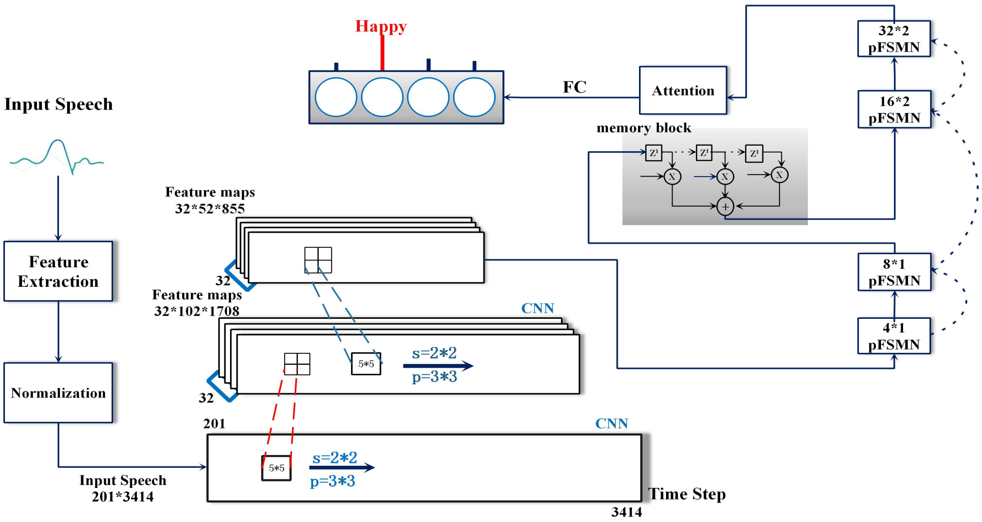

Figure 3: *CNN-pFSMN-Attention structure, in which s=2*2 denotes CNN strides, p=3*3 denotes CNN padding size. The time orders and strides are from 4 to 32 and 1 to 2 respectively, which forms a pyramid FSMN. Timestep attention mechanism is added after the last pFSMN layer to automatically extract useful information from feature map. FC denotes fully connection between attention and output layer.*

various useful information while focusing on different region. We will shift attention to clothes and hair if we want to know a person's gender, for example.

With attention mechanism, each element in the output sequence is conditional on selective elements in the input sequence. This increases the computational burden of the model, but results in a more accurate and better performing model. In most implementations, attention is realized as a weight vector (often as the output of a softmax function), the dimension of which is equal to the length of the input sequence. In our work, a piece of speech is divided into many small segments, which are called timestep in neural network. It is obvious that not every timestep is useful for SER task while a speech contains a lot of white space, hence the model needs to focus on some certain region. Based on this, we propose a timestep attention mechanism as shown in Figure 2. It can be described as the following formulations:

$$a_t = Average(\boldsymbol{h_t}) \tag{5}$$

$$\boldsymbol{s} = \mathcal{S}(\boldsymbol{W_2}(f(\boldsymbol{W_1 a} + \boldsymbol{b_1})) + \boldsymbol{b_2}) \tag{6}$$

$$\boldsymbol{y} = \boldsymbol{Xs} \tag{7}$$

$\boldsymbol{h_t}$ denotes the last hidden layer's output of t-th timestep, we use the average to represent the whole timestep which denoted by a_t. And $\boldsymbol{a}$ denotes the vector consists of a_t in all timesteps. $f(\cdot)$ is the activation function (sigmoid or ReLU), we just use ReLU in our work. $\mathcal{S}(\cdot)$ denotes the Softmax function to calculate the weight of each timestep in final output. And $\boldsymbol{X}$ denotes the input of timestep attention mechanism.

3.3. CNN

As shown in Figure 3, instead of directly using DFSMN layers, a 2-layer CNN module is applied at the front end. We just use speech spectrograms as the input feature, which could be fed into 2D-CNN directly as a map. This is inspired by image-processing. In speech emotion recognition tasks, the frequency and timesteps correspond to pixel values of x and y respectively. Besides, subsampling is applied to extract more robust features and reduce the features' size with the network goes deeper. It improves accuracy significantly.

4. Experiments

4.1. Experiment Setup

We evaluate our SER model using IEMOCAP [22] corpus, which is a database where two actors communicate in each session to elicit specific type of emotions. The utterances were segmented and with one categorical label, which is among angry, fear, excited, neutral, disgust, surprised, sad, happy, frustrated, other and XXX. XXX was the case that the annotators were not able to have agreement on the label. In this work, we only select 5 classes: angry, excited, happy, neutral and sad. The total number of utterances used is 5531. Happy and excited emotions were combined as happy in order to balance the number of samples in each emotion class. Additionally, we randomly selected 10% of the total data as testing subjects. The rest data were used as training data. And 10% of the training data was used as validation data to check whether we need early stopping. The data sizes in our experiments are listed in Table 1.

Table 1: *Number of utterances in IEMOCAP corpus*

	Angry	Happy	Neutral	sad
train	886	1323	1375	886
val	113	153	157	108
test	104	160	176	90

The corpus has video and audio channels while we only used audios in this study. The audio was collected by high quality microphones (Schoeps CMIT 5U) at the sample rate of 48 kHz. We downsampled them to 16 kHz and extracted a 201D acoustic feature. Different from other paper, we just use the spectrograms as input. The extraction was performed within a 25 ms window whose shifting step size was 10 ms (100 fps). The acoustic feature sequence was z-normalized within each ut-

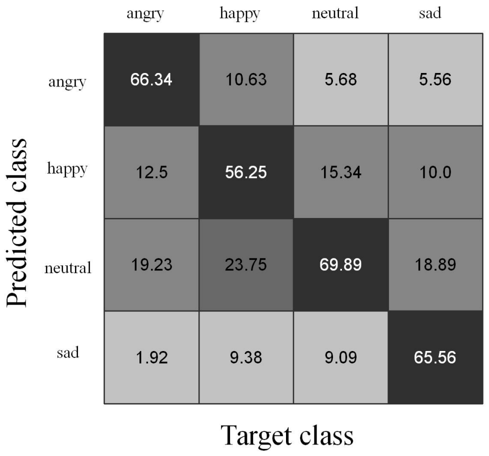

Figure 4: *Error distribution of the C-pFSMN-A model.*

terance.

The network architecture is shown in Figure 3. Two 5*5 conv layers are employed at the front. 4 FSMN blocks' hidden layer and memory block have 256 and 128 nodes respectively. Both CNN and pFSMN layers are followed by batch normalization layer to avoid overfitting. The time orders and strides are from 4 to 32 and 1 to 2 respectively. And the training of our model is performed based on the SGD optimizer using pytorch, and the batch size is set to 32.

4.2. Results and Discussion

To measure performance of systems, we report overall accuracy on test examples (weighted accuracy, WA) and average recall over different emotion categories (unweighted accuracy, UA) in addition to recall in each class. The performance of several comparative system and the proposed systems on IEMOCAP corpus are listed in Table 2, in which C denotes CNN layers and A denotes Attention Mechanism.

The results show that the pFSMN improved by 2.47% abosolute improvement in UA over LSTM, which proves that FSMN has even better sequence model performance. HSF-CRNN [23] is an imporved CNN combined RNN approach, proposed by Luo, which uses hand-crafted speech features as input, and our model achieves 0.53% and 3.99% absolute improvement in UA and WA respectively. It proves that we could automatically extract useful information from spectrograms without help of commonly used hand-crafted speech features. We also built a basic CNN-biLSTM model for comparison. Its accuracy of sad samples is better than our approach while rest are much worse. To show that the attention mechanism works, we built a C-pFSMN model without attention while rest is exactly the same. The result presents that the proposed attention mechanism performs well in SER task, which achieve 6.3% absolute improvement in UA compared with C-pFSMN. In addition, front CNN layers extract more sophisticated features to improve model performance as expected.

The C-biLSTM in the table 2 was built with 2-CNN layers and 2-Bi-LSTM which had 256 nodes in hidden layers. It's similar to our model and widely used in sequence modeling tasks. Hence we also compared the computing resources of C-LSTM with our approach, which measured by the number of parameters and training time required. As presented in table 3, our model only has 1.85M parameters and 64-min of training, which much faster than C-LSTM model. That means we could achieve better performance while requiring less computing resources.

Table 2: *Performance(%) of different approaches on IEMOCAP dataset*

Approach	**angry**	**happy**	**sad**	**neutral**	**WA**	**UA**
LSTM[24]	52.47	43.22	62.40	54.8	53.43	53.23
HSF-CRNN[23]	/	/	/	/	60.35	63.98
C-biLSTM	63.46	47.5	**73.33**	63.06	60.19	61.84
pFSMN	53.84	53.75	56.67	58.52	55.85	55.70
C-pFSMN	57.70	52.50	54.44	68.18	59.06	58.21
C-pFSMN-A	**66.34**	**56.25**	65.56	**69.89**	**64.34**	**64.51**

Table 3: *Computing resources comparison*

Approach	**Parameters(M)**	**Time(min)**
C-biLSTM	2.12	113
C-pFSMN-A	**1.85**	**64**

We also analyze the predictions of the C-pFSMN-A. Figure 4 shows the confusion matrices. The most striking observation is that the model predicts neutral for 23.75% happy samples. This mistake becomes more plausible with our data preprocessing. In our work, happy and excited emotions were combined as happy in order to balance the number of samples in each emotion class. We think this operation makes it somewhat harder for the model to predict the happy samples, although it improves the overall performance of the model. The other categories have higher recognition accuracy benefits from the preprocessing.

5. Conclusion

In this work, we proposed a CNN-pFSMN based architecture with timestep attention mechanism to improve emotion recognition from speech. By applying CNN, pyramidal memory block structure and timestep attention, we achieved 2.67% absolute improvement on the IEMOCAP dataset compared to commonly used conventional CNN-biLSTM model which required much more computing resource. We proved that FSMN has strong sequence model ability even better than LSTM in the SER task. Instead of well-established hand-crafted speech features, we just choose spectrograms as the input of model, which reduces the impact of human factors.

In the future, we plan to explore more effective ways of combining information from different modalities while human emotion is affected by many factors such as gender, speaker, lanuage and recording environment variations. We are prepared to apply our model with multi-task learning technology and explore more efficient fusion features in the following research.

6. Acknowledgements

The research was partly supported by Beijing Natural Science Foundation(4194084), China Postdoctoral Science Foundation(2018M641199) and Fundamental Research Funds for the Central Universities (FRF-TP-18-060A1).

7. References

[1] D. Neiberg, K. Elenius, and K. Laskowski, "Emotion recognition in spontaneous speech using gmms," in *Ninth International Conference on Spoken Language Processing*, 2006.

[2] K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, and A. Aucinas, "Emotionsense: a mobile phones based adaptive platform for experimental social psychology research," in *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, 2010, pp. 281–290.

[3] K. W. Gamage, V. Sethu, P. N. Le, and E. Ambikairajah, "An i-vector gplda system for speech based emotion recognition," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2015, pp. 289–292.

[4] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[5] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 2. IEEE, 2003, pp. II–1.

[6] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll, "Emotion recognition in the noise applying large acoustic feature sets," in *Proc. Speech Prosody 2006, Dresden*, 2006.

[7] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech communication*, vol. 52, no. 7-8, pp. 613–625, 2010.

[8] N. Yang, J. Yuan, Y. Zhou, I. Demirkol, Z. Duan, W. Heinzelman, and M. Sturge-Apple, "Enhanced multiclass svm with thresholding fusion for speech-based emotion classification," *International journal of speech technology*, vol. 20, no. 1, pp. 27–41, 2017.

[9] F. Tao and G. Liu, "Advanced lstm: A study about better time dependency modeling in emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 2906–2910.

[10] P.-W. Hsiao and C.-P. Chen, "Effective attention mechanism in dynamic models for speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2526–2530.

[11] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 1537–1540. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2015/i15_1537.html

[12] S. E. Eskimez, Z. Duan, and W. Heinzelman, "Unsupervised learning approach to feature analysis for automatic speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5099–5103.

[13] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller, "Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with lstm neural networks," in *17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016)*, 2016, pp. 3593–3597.

[14] P. J. Werbos *et al.*, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[15] Y. Bengio, P. Simard, P. Frasconi *et al.*, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[16] S. Zhang, C. Liu, H. Jiang, S. Wei, L. Dai, and Y. Hu, "Feedforward sequential memory networks: A new structure to learn long-term dependency," *arXiv preprint arXiv:1512.08301*, 2015.

[17] X. Yang, J. Li, and X. Zhou, "A novel pyramidal-fsmn architecture with lattice-free mmi for speech recognition," *arXiv preprint arXiv:1810.11352*, 2018.

[18] S. Zhang, M. Lei, Z. Yan, and L. Dai, "Deep-fsmn for large vocabulary continuous speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5869–5873.

[19] R. Van Bezooijen, S. A. Otto, and T. A. Heenan, "Recognition of vocal expressions of emotion: A three-nation study to identify universal characteristics," *Journal of Cross-Cultural Psychology*, vol. 14, no. 4, pp. 387–406, 1983.

[20] R. W. Picard, *Affective computing*. MIT press, 2000.

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

[22] 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.

[23] D. Luo, Y. Zou, and D. Huang, "Investigation on joint representation learning for robust feature extraction in speech emotion recognition," in *Proc. Interspeech 2018*, 2018, pp. 152–156. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1832

[24] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," in *Proc. Interspeech 2018*, 2018, pp. 247–251. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-2466