Speech Emotion Recognition via Ensembling Neural Networks

Danqing Luo[1,3], Yuexian Zou[1*], Dongyan Huang[2*]
[1]ADSPLAB/ELIP, School of ECE, Peking University, Shenzhen, China
[2] Human Language Technology, Institute for Infocomm Research/A*STAR, Singapore
[3]IMSL Shenzhen Key Lab, PKU-HKUST Shenzhen Hong Kong Institution
*E-mail: zouyx@pkusz.edu.cn, huang@i2r.a-star.edu.sg

***Abstract*—Deep Neural Network (DNN) based speech emotion recognition (SER) methods have demonstrated competitive performance compared to traditional SER approaches. However, from literatures, it can be seen that the confusion matrices of different SER methods varied a lot, which indicates that different DNN architecture has different capability of modeling different emotion cues from speech. It also means that single classifier hardly performs well on all speech emotion categories, which may be possibly due to data imbalance and the limitation of classifier. Motivated by the improved research results of ensemble learning, this paper investigates an ensemble method for SER via aggregating results from several base classifiers. In this study, considering the outstanding performance of Recurrent Neural Network (RNN) in different speech tasks and Residual network (ResNet) in image related classification, we chose RNN and ResNet acting as base classifiers. Experiments show that our proposed ensemble SER system outperforms the state-of-art single classifier-based SER system.**

***Keywords*—Speech emotion recognition; recurrent neural network; residual network; ensemble learning.**

I. INTRODUCTION

Emotion recognition is an important ability for computer to provide good experience in human-computer interaction. Due to this, speech emotion recognition (SER) has become an active research area with variety applications. Literature shows that there are many research outcomes of SER since 1980s [1]. However, in our opinion, SER is indeed a difficult machine hearing task. Compared with speech recognition task, there are relatively few public research datasets for SER task. As a result, up to now, SER is still an open challenge in research community.

In principle, SER system derives an emotion recognition decision on either short segment or utterance level, depending on the way target label is assigned. For segment-level emotion recognition, the target label is assigned to each single segment of the utterance. Low-level descriptors (LLDs) are extracted from speech frame as input to the sequential classifier, which mainly uses Gaussian mixture model (GMM) [2] or hidden Markov model (HMM) [3] to model the distribution of emotion state of speaker. For utterance-level emotion recognition, the target label is given on a whole utterance. Statistics are computed for LLDs over all frames in an utterance as global feature, which is input to a discriminative classifier such as support vector machine (SVM) [4], decision trees [5] or K-nearest neighbor [6]. Research results indicate that feature design is an important step for conventional SER methods. However, it is not a trivial job to figure out the optimal feature set for SER tasks. Usually, researchers have to select them empirically through experimental trails.

Recently, deep neural network has been introduced to the field of SER [6-9] since it has the ability to learn hierarchical high-level representation from raw features. In [7], DNN is designed to learn short time segment-level features while the extreme learning machine (ELM) is used for final utterance-level classification. In [8], authors proposed a Recurrent Neural Network (RNN) based framework with maximum-likelihood learning criterion to model random label sequence of an utterance, which gives an improved SER recognition accuracy. Work in [9] investigated different RNN architectures and introduced a local attention mechanism to weight frames accordingly. In [10], the convolutional neural network is successfully designed to learn salient and discriminative features for SER.

Carefully evaluating the SER results from different methods discussed above but on same dataset, there is an interesting finding from the confusion matrix, where it is clear that different classifier with same LLDs can gain different recognition accuracy on each emotion category. This indicates that single SER classifier hardly performs well on all emotion categories. For example, a SVM-based classifier may fail to correctly recognize the happy emotion while DNN-based one is superior in that. Such different capability is possibly related to data imbalance and SER system's modeling capability.

In this study, triggered by this finding, aiming at improving SER recognition ability, we propose an ensemble learning SER method with two neural network base classifiers. As suggested in [11], in order to obtain better generalization, base classifiers used in ensemble system are required to be as much different as possible in terms of architecture. Hence, in this work, RNN with LSTM block and wide residual network are chosen as base classifiers since the former has proved suitable for processing sequential data and the latter gives the state-of-art in image classification. Besides, to balance the short term characterization at frame level and long term global aggregation at utterance level, we design the RNN to be trained at utterance-level for capturing contextual emotion information and the wide residual network at segment-level for modeling the subtle emotion cues, respectively.

The paper is organized as follows. In the next section, RNN and wide residual network are introduced in brief for presentation clarity. The proposed ensemble SER system and its subsystems are given in section 3. Experiments will be shown in section 4 and conclusion is drawn in section 5.

II. PRELIMINARY

Two neural networks are introduced as the base classifiers of our proposed ensemble SER system, namely recurrent neural network and wide residual network. For presentation completeness, here we briefly introduce the basic concepts.

A. RNN with LSTM block

RNN is powerful for processing sequential data due to the unique net framework. Recurrent connection of its hidden unit propagates hidden state from a time step of input sequence to the next time step. Such recurrence repeats from the first time step to the last, that the sequential correlation can be well modeled.

Research shows that RNN's performance degrades when input a long sequence because of the gradient vanishing problem. To overcome it, LSTM (Long Short Term Memory) model with a new structure is proposed [12].

Generally, a LSTM block contains four main elements: an input gate *i*, a forget gate *f*, an output gate *o* and a self-connected memory cell *c*. These gates are responsible for modulating the interaction between memory cell's state of current time step and its state of previous time step. For input x_t at time step *t*, let i_t, f_t, o_t, c_t and h_t denote the value of i, f, o gate, memory cell and LSTM layer output, respectively, which are given by

$$i_t = \tanh(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (1)$$

$$f_t = \tanh(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (2)$$

$$c_t = i_t \cdot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) + f_t \cdot c_{t-1} \quad (3)$$

$$o_t = \tanh(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_0) \quad (4)$$

$$h_t = o_t \cdot \tanh(c_t) \quad (5)$$

where $W_{x.}$, $W_{h.}$, $W_{c.}$ represents the weight matrix between layer input, layer output, memory cell and each gate respectively; $b_.$ represents bias for each gate.

B. Wide Residual Network

It is a common sense that CNN gets more difficult to train as it goes deeper due to gradient vanishing problem. To overcome this, a residual network (ResNet) is proposed [13], which has shown great success when its depth is much deeper than conventional CNN while the improved recognition accuracy is achieved.

Inspired by the success of ResNet going deeper, an alternative approach called wide residual network (WRN) was proposed [14], which achieves better accuracy with decreased depth but increased width in network architecture.

In principle, ResNet is sequentially stacked by residual block, which usually contains two consecutive convolutions with batch normalization (BN) and ReLu activation preceding each. Compared to conventional ResNet, WRN increases the number of filters in each convolution by *k* times to widen the convolutional layers and enhance the representation ability of each layer consequently. Research study demonstrates that WRN with widened structure can get the same accuracy as conventional ResNet with much shallower depth [14]. Fig. 1 shows the block diagram of a residual block and the structure of a WRN. Four types of residual block are used in WRN. All of them has a convolution filter with size of 3×3 and their filter number are $16, 16\times k, 32\times k, 64\times k$ respectively. Consecutive *N* residual blocks of the same type is viewed as a group. These four groups are stacked on end, and finally the average pooling layer and softmax layer are added to complete a WRN.

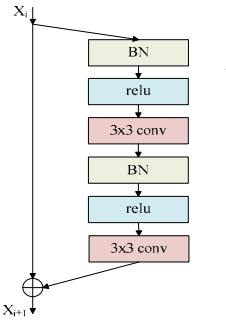

Fig. 1 (a) residual block

Group name	Group architecture
Cov 1	[3×3 16]
Cov 2	$\begin{bmatrix}3\times3 & 16\times k\\ 3\times3 & 16\times k\end{bmatrix}\times N$
Cov 3	$\begin{bmatrix}3\times3 & 32\times k\\ 3\times3 & 32\times k\end{bmatrix}\times N$
Cov 4	$\begin{bmatrix}3\times3 & 64\times k\\ 3\times3 & 64\times k\end{bmatrix}\times N$
Avg-pool	[8 8]

Fig. 1 (b) structure of WRN

III. PROPOSED SER METHOD

As discussed above, we propose an ensemble method for SER via aggregating results from several base classifiers. In detail, RNN and WRN are taken as base classifiers, among which WRN is introduced to the field of SER for the first time. Each base classifier is viewed as a subsystem, which can implement SER independently. In this section, we will introduce the details of two subsystems and our proposed ensemble SER system.

A. RNN-based subsystem

The block diagram of RNN-based subsystem is shown in Fig. 2. It is noted that the input to RNN is time sequential feature vectors of an utterance, which is denoted as s(1) ,…, s(*T*). Actually, s(*t*) is the *t*th segment extracted by a fixed window and it is taken as the input of RNN-based subsystem at the *t*th time step. The target label is given on a whole utterance. As illustrated in Fig. 2, at each time step, the raw feature vector is processed by a dense layer and

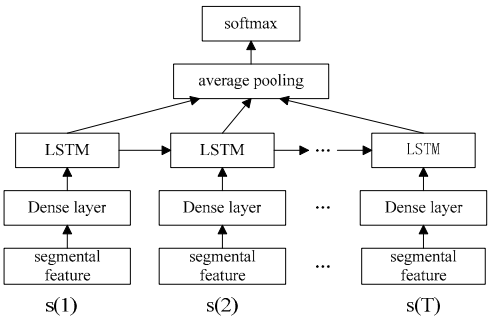

Fig. 2 RNN-based subsystem

then a LSTM layer. Outputs from different time step pool together and the average is computed as the utterance's global feature, which is fed to a softmax layer and produce a probability distribution vector over emotion categories. The cross entropy is taken as loss function when training the network.

Moreover, to better understand the formation of segment input, one example is given here. Given feature vector f(t) extracted from frame time t and window size w, the segmental feature vector s(t) can be presented as [f(t-w),…, f(t),…, f(t+w)]. In this work, f(t) consists of 12-d MFCC, energy, zero crossing rate, voice probability, fundamental frequency and their delta features across time frames, 32 dimensions in total.

B. *WRN-based subsystem*

The block diagram of WRN-based subsystem is shown in Fig. 3. Spectrogram of an utterance is first divided into segments and then a WRN is trained on segmental spectrogram to generate emotion category distribution for each segment. From these segment-level emotion category distributions, utterance-level global features are constructed and input to a softmax classifier to determine the category of the whole utterance. To conclude, this subsystem consists of two parts, a WRN classifying on segment-level and a softmax classifier on utterance-level.

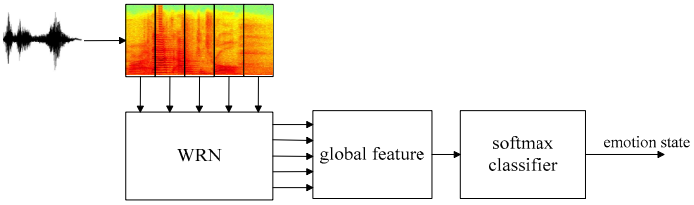

Fig. 3 WRN-based subsystem

In training stage, WRN is firstly trained and target label for each segment is assigned to the emotion category of utterance it belongs to. Then global features of training utterances are computed from the output of WRN, which are used to train the softmax classifier in the end with their utterance label. The details of global feature are given next.

Global features are a series of statistics. Let there are K emotion categories to be recognized and $P_s(E_k)$ denote the probability of segment s belonging to emotion E_k. Taking the ith utterance for example, statistics are computed for $k = 1, 2, \ldots, K$ using following equations, where U denotes the set of segments in utterance i:

$$f_k^1 = \sum_{s \in U} P_s(E_k) / |U| \quad (6)$$

$$f_k^2 = \min P_s(E_k), s \in U \quad (7)$$

$$f_k^3 = \max P_s(E_k), s \in U \quad (8)$$

$$f_k^4 = |P_s(E_k) > 0.5| / |U| \quad (9)$$

where f_k^1, f_k^2 and f_k^3 represent the mean, minimum and maximum probability of E_k for utterance i respectively. f_k^4 gets the fraction of segments whose probability is larger than 0.5. Concatenating statistics of all K categories together into one vector, the global feature vector is formed as $[f_1^1, f_1^2, f_1^3, f_1^4, ..., f_K^1, f_K^2, f_K^3, f_K^4]$.

C. *Ensemble system*

RNN-based and WRN-based subsystems both generate probability distribution vector to determine the recognition result. To realize ensemble, we sum up two vectors to form a new global feature. Specifically, given training dataset $(x_i, y_i), i = 1, ..., N$, where x_i is a raw speech signal, y_i is the target class and N is the number of training sample, firstly RNN and WRN-based subsystems are trained independently. For the ith utterance, each subsystem outputs a probability vector denoted as $(r_i^1, ..., r_i^K)$ and $(w_i^1, ..., w_i^K)$ respectively. Therefore, a new global feature vector P_i is given by element-wise addition expressed as

$$P_i = (p_i^1, ..., p_i^K) = (r_i^1, ..., r_i^K) + (w_i^1, ..., w_i^K) \quad (10)$$

In ensemble layer, $(P_i, y_i), i = 1, ..., N$ is used as training data to train a softmax classifier.

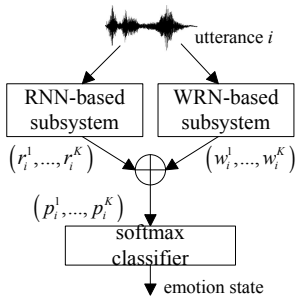

Fig. 4 Ensemble system

In testing stage, for a test utterance, the probability distributions from RNN and WRN-based subsystems are computed first. Then, the global feature vector P_i is generated by (10). Finally softmax yields the prediction.

IV. EXPERIMENTS AND ANALYSIS

A. *Experimental Setup*

The performance of our proposed ensemble SER system is evaluated using the public research corpus IEMOCAP. The corpus is composed of five sessions from 10 speakers, and in each session a pair of actors talk to each other according to given scripts or improvise in a pre-defined situation. In our experiments, we take four emotion categories: angry, neutral, sadness and excitement, which represent the majority of emotion categories in the corpus. Experiments are conducted in a speaker-independent manner. Four sessions from 8 speakers are chosen as training data, the remaining one from 2 other speakers as test. The detail of used dataset is described in Table I. It is clear that training samples are pretty imbalanced across all categories.

For RNN-based subsystem, speech signal is converted into frames with a sliding window of 25 ms at step of 10 ms. A segment is formed by 25 successive frames without overlap, thus the length of one segment is $10\text{ms} \times 24 + 25\text{ms} = 265\text{ms}$.

TABLE I
DATASET DESCRIPTION

Emotion category	Training set (utterances)	Test set (utterances)	Total (utterances)
angry	933	170	1103
neutral	1324	384	1708
sad	839	245	1084
excitement	742	299	1041

Research has shown that a speech segment longer than 250ms can encode sufficient emotion information [15]. In this way, 32-dim low level feature is expanded to 800-dim segmental feature vector. Then all segments are normalized into zero mean and unit variance. Therefore, the input layer of RNN has 800 neurons corresponding to the 800-dim segmental feature vector. The sigmoid dense layer and LSTM layer both have 256 nodes.

For WRN-based subsystem, the frame width is 32 ms and sliding step is 10 ms. 512-point short time Fourier transformation (STFT) is applied to each Hanning windowed frame. We use a log-scale magnitude of STFT to extract spectrogram. Size of one segment is set to 30 frames, thus the length of a segment is 322 ms. Every two adjacent segments has an overlap of 15 frames. As a result, input of WRN is a data matrix with 257×30 dimensions. Total depth of WRN is 28, with N=4 and k=2.

To measure the performance, the weighted accuracy (WA) and the unweighted accuracy (UA) are reported for two subsystems and the ensemble system, where WA is the classification accuracy on the entire dataset and UA is the average accuracy of each category.

B. *Experimental Results*

Table II shows the confusion matrices directly resulted from two subsystems. We can see that neutral emotion is the easiest to distinguish, angry second, sadness the next and excitement is the most difficult to recognize, which is the same as decreasing order of each emotion category's training sample number.

TABLE II
CONFUSION MATRIX OF SYSTEMS

	Ang	Neu	Sad	Exc
Ang	61.8	15.3	10.0	12.9
Neu	10.4	70.0	15.4	4.2
Sad	6.5	28.2	60.8	4.5
Exc	20.4	36.8	9.7	33.1

(a) RNN-based subsystem

	Ang	Neu	Sad	Exc
Ang	74.7	18.2	2.4	4.7
Neu	4.9	76.6	17.7	0.8
Sad	5.3	34.3	60.0	0.4
Exc	26.1	57.2	3.3	13.4

(b) WRN-based subsystem

	Ang	Neu	Sad	Exc
Ang	67.1	20.0	3.5	9.4
Neu	5.2	77.6	14.1	3.1
Sad	5.3	31.8	58.4	4.5
Exc	17.1	45.2	5.3	32.4

(c) Ensembel system

More specifically, WRN-based subsystem has the ability to recognize angry and neutral better than RNN-based one and they perform equally on sadness, but RNN-based subsystem is more sensitive to excitement. On the whole, RNN-based subsystem can evenly recognize each emotion better over an imbalanced data set. Moreover, in Table II, the confusion matrix from ensemble system is presented and we can see that the ensemble system obtains recognition ability for angry and neutral from WRN-based subsystem, and that for excitement from RNN-based subsystem while maintaining the same level of recognition ability for sadness as two subsystems. To judge performance on the whole dataset, the UA and WA of each system are presented in Table III. Compared to RNN-based SER method and WRN-based one, the proposed ensemble system shows around 2% and 3% improvement in UA and WA, respectively. Moreover, on the same dataset, comparing the results of 0.5213 in UA and 0.5791 in WA using classical DNN-ELM method reported by [8], our proposed method exceeds slightly.

TABLE III
COMPARISON OF DIFFERENT SYSTEMS IN UA AND WA

system	UA	WA
RNN-based	0.5644	0.5665
WRN-based	0.5616	0.5537
ensemble	0.5887	0.5938

V. CONCLUSION AND FUTURE WORK

In this paper, we developed an ensemble SER method using RNN and WRN as base classifiers. This method aims at improving SER accuracy by combining advantages of two DNN classifiers with different network architectures. Specifically, RNN models the sequential information and predicts at utterance level, while WRN learns the representation of spectrogram and predict at segment level. Experiments confirm the effectiveness of the ensemble SER method compared with non-ensemble SER method. Besides, we initially employ WRN in the field of SER and it shows comparable with mainstream RNN-based method. It is a bit disappointed that the improvement from ensemble learning approach is not significant in this study. There are possible two reasons, one is the data imbalance problem and another is the base classifier design problem. In our future work, we will further investigate the data augmentation method to relieve the influence of data imbalance and introduce bi-directional LSTM to reinforce the system's modeling ability for speech sequence.

ACKNOWLEDGMENT

This project was partially supported by Shenzhen Science & Technology Fundamental Research Programs (No: JCYJ20170306165153653 & JCYJ20150331165212372).

REFERENCES

[1] Ververidis, Dimitrios, and Constantine Kotropoulos. "Emotional speech recognition: Resources, features, and methods." Speech communication 48.9 (2006): 1162-1181.

[2] Neiberg, Daniel, Kjell Elenius, and Kornel Laskowski. "Emotion recognition in spontaneous speech using GMMs." Interspeech. 2006.
[3] Nwe, Tin Lay, Say Wei Foo, and Liyanage C. De Silva. "Speech emotion recognition using hidden Markov models." Speech communication 41.4 (2003): 603-623.
[4] Mower, Emily, Maja J. Mataric, and Shrikanth Narayanan. "A framework for automatic human emotion classification using emotion profiles." IEEE Transactions on Audio, Speech, and Language Processing 19.5 (2011): 1057-1070.
[5] Kostoulas, T. P., and Nikos Fakotakis. "A speaker dependent emotion recognition framework." Proc. 5th International Symposium, Communication Systems, Networks and Digital Signal Processing (CSNDSP), University of Patras. 2006.
[6] Pao, Tsang-Long, et al. "Comparison of several classifiers for emotion recognition from noisy mandarin speech." Intelligent Information Hiding and Multimedia Signal Processing, 2007. IIHMSP 2007. Third International Conference on. Vol. 1. IEEE, 2007.
[7] Han, Kun, Dong Yu, and Ivan Tashev. "Speech emotion recognition using deep neural network and extreme learning machine." Interspeech. 2014.
[8] Lee, Jinkyu, and Ivan Tashev. "High-level feature representation using recurrent neural network for speech emotion recognition." INTERSPEECH. 2015.
[9] Mirsamadi, Seyedmahdad, Emad Barsoum, and Cha Zhang. "AUTOMATIC SPEECH EMOTION RECOGNITION USING RECURRENT NEURAL NETWORKS WITH LOCAL ATTENTION."
[10] Mao, Qirong, et al. "Learning salient features for speech emotion recognition using convolutional neural networks." IEEE Transactions on Multimedia 16.8 (2014): 2203-2213.
[11] Cunningham, Padraig, and John Carney. "Diversity versus quality in classification ensembles based on feature selection." European Conference on Machine Learning. Springer Berlin Heidelberg, 2000.
[12] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.
[13] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
[14] Zagoruyko, Sergey, and Nikos Komodakis. "Wide residual networks." arXiv preprint arXiv:1605.07146 (2016).
[15] Kim, Yelin, and Emily Mower Provost. "Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions." Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.