JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 36, 561-576 (2020)
DOI: 10.6688/JISE.202005_36(3).0006

A Robust Equalization Feature for Language Recognition

WEN-JIE SONG, CHEN CHEN, TIAN-YANG SUN AND WEI WANG
School of Computer Science and Technology
Harbin Institute of Technology
Heilongjiang, 150001 P.R. China
E-mail: {64228128; 794452272}@qq.com; sun_tianyang@yahoo.com; wangwei_hitwh@126.com

The performance of language recognition system is mainly determined by feature extraction and model training. In this paper, a robust equalization feature for language recognition is proposed, which utilizes the common features of the speech spectrum mean vector to calculate a global mean vector. The spectrum mean vector of each segment is equalized on the global mean vector, and the equalization features are obtained. In model training, Gated Recurrent Unit (GRU) of Recurrent Neural Network (RNN) is applied to language recognition, in which GRU can reduce the amount of computation and shorten the training time. The experimental results show that the proposed method outperforms the baseline system on the NIST LRE 2007 corpus.

Keywords: language recognition, deep neural network, gated recurrent unit, feature extraction, model training

1. INTRODUCTION

Language recognition is a branch of speech recognition technology, which is used to judge the language of the speech segments. Usually, there are three phases for recognition system to analyze speech segments, including feature extraction, model training and model testing. In the feature extraction phase, the speech data of different languages is converted into vector sequences. In the model training phase, a reference model is established by those extracted vector sequences according to a certain training algorithm. In the model testing phase, the test feature vectors are compared with each reference model, and the recognition result is determined according to the likelihood.

In feature extraction, short-term cepstral features such as Mel-frequency cepstral coefficients (MFCC) has been the principal features in language recognition field [1]. MFCC simulates the filter property of cochlear system, which is proved to be practical for improving the performance. Although MFCC achieves desirable performance in ideal condition, the result would be rapidly deteriorated in noisy situation [1]. To compensate for the influence of noisy situation, the researchers put forward different robust characteristics, such as Cepstral Mean Subtraction (CMS) [2, 3], Cepstral Mean and Variance Normalization (CMVN) [4, 5] and RelAtive SpecTrAl (RASTA) [6, 7]. CMS assumes that the convolution noise of channel distortion is linear and time invariant. Therefore, in the spectrum Mel domain, the mean of the frame can be subtracted with the current frame to eliminate the influence of the time invariant channel, but it can not eliminate the distortion of the annex channel. Similarly, CMVN simultaneously regularizes the mean and variance of

Received October 8, 2018; revised December 30, 2018; accepted February 14, 2019.
Communicated by Berlin Chen.

speech features, which eliminates the influence of time invariant channels and additive noise. Rasta filtering technology can eliminate lower and higher modulation frequency to reduce the channel distortion. Although the methods above can achieve better results in channel distortion and relatively stable additive noise environment, the effect is not ideal under the condition of non-stationary noise [8]. In this paper, we propose the spectrum equalization feature to deal with non-stationary noise. In frequency domain, the spectrum of each speech segment is averaged on time axis to obtain the spectrum mean vector. It is found that these mean vectors are locally different, but the overall distribution has a same trend. To find the common characteristics of these mean vectors, we average all the mean vectors to get the global mean vector. With the principle of histogram equalization in image area, each mean vector is equalized on the global mean vector to make it more similar to the global one, which weakens the local jitter and obtains clearer and more intuitional spectral feature parameters with less noise effects.

In the model training, deep learning [9-11] is a famous method focusing on the field of speech signal processing. Deep Neural Network (DNN) learns statistical rules from a large number of training samples, and then predicts unknown events. Compared with the system based on artificial rules in the past, it shows superiority in many aspects. However, the data flow in normal fully connected neural network is directed, from the input layer to the hidden layer, and finally to the output layer, and the nodes in the same hidden layer are not connected. Such network structure is unable to solve many sequential problems, resulting in the generation of Recurrent Neural Network (RNN) [12-14]. RNN can memorize the previous information and apply it to the current output. The nodes in the same layer are no longer unconnected and the input of the hidden layer includes not only the output of the input layer but also the output of the previous hidden layer. Theoretically, RNN can process sequence data with arbitrary length, while in practice, excessively long sequence of data often brings vanishing gradient. To solve this problem, researchers proposed Long Short-Term Memory (LSTM) [15-17], which is a variant of RNN. It controls the forward data flow and backward data flow by setting a block structure that includes three memory gates (input gate, output gate, and oblivion gate) which solve the vanishing gradient effectively. Gated Recurrent Unit (GRU) [18, 19] is a variant of LSTM which can preserve the excellent performance of LSTM and simplify the block structure, reducing the amount of computation and shortening the training time.

2. EQUALIZATION FEATURE EXTRACTION

The process of traditional MFCC feature extraction is shown in Fig. 1. Firstly, the preprocessed speech is framed into short frames, and Fast Fourier Transform (FFT) is calculated to obtain spectrum. Secondly, the Mel Triangle Filter bank is applied to the power spectra and the energy in each filter is summed up. Then, the logarithm of all filter bank energies is taken. Finally, Discrete Cosine Transform (DCT) is performed on the log filter bank energies to get MFCC. The first-order and second-order delta of the MFCC are computed as the dynamic features and stitched together with the static features to form the features of each frame.

In order to prevent the amplitude of the spectrum from being concentrated on a certain frequency domain and being suppressed on other frequency domains, the spectrum needs to be equalized so that every ingredient can be expressed clearly, which contribute to the

generation of equalization feature extraction. After the spectrum is obtained by FFT, spectral normalization, spectral averaging and spectral equalization are added into the process, as shown in Fig. 2.

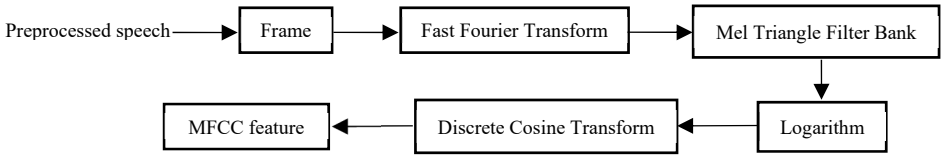


Fig. 1. The process of traditional MFCC feature extraction.

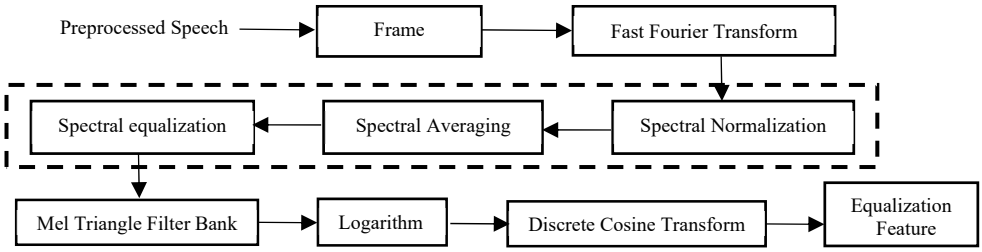


Fig. 2. The process of equalization feature extraction.

We assume that there are K voiceprint spectrums corresponding to K speech segments, and the kth voiceprint spectrum is an $N \times M$ matrix (where N is the dimension of the spectral feature and M is the number of frames), $A^k = (x_1, x_2, \ldots, x_M)$, where $x_m \in R^N (m = 1, 2, \ldots, M)$ is the vector of each frame. Each feature vector is normalized as:

$$||x_m||_2^2 = 1. \tag{1}$$

Different speech segments have different amplitude distributions on different frequencies in the spectrums. To observe the distribution, the voiceprint spectrum of each segment is averaged over the time axis to obtain the spectrum mean vector. The mean vector of the kth voiceprint spectrum is calculated as:

$$v_k = \left(\sum_{m=1}^{M} x_m\right) \Big/ M. \tag{2}$$

The mean vector distribution of some speech segments is shown in Fig. 3.

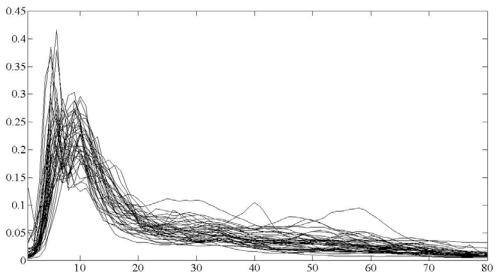


Fig. 3. The mean vector distribution for some speech segments.

Fig. 3 illustrates that although the distribution of the spectrum mean vectors of these speech segments in different languages is locally different, the overall trend is similar. To observe the trend, we average the k mean vectors to get the global mean vector.

$$\overline{V} = (\sum_{k=1}^{K} v_k) \Big/ K \tag{3}$$

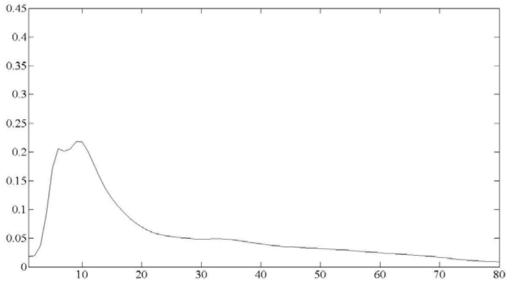


Fig. 4. The distribution of global mean vector.

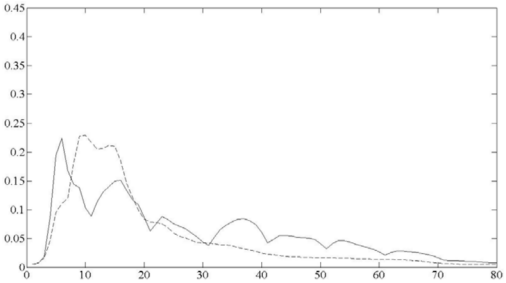


Fig. 5. Two spectrum mean vectors of "Thai".

In Fig. 4, the distribution of the global mean vector shows a remarkable peak on the voice print spectrum within in a certain range of frequency, and it indicates that the amplitude distribution in this frequency interval is the most prominent. Thus, this frequency interval can be regarded as the main frequency domain of the speech. At both ends of the main frequency domain, the amplitude distribution gradually decreases. However, in practical application, due to the differences in external factors like the channel environment, the amplitude distribution of the same language will change, as shown in Fig. 5. In Fig. 5, the dotted line indicates the mean vector corresponding to a speech segment with a relatively clean background, while the solid line represents the mean vector with the channel noise.

In order to suppress the noise influence, the ratio c_k of the corresponding elements in $\overline{V}$ and v_k is calculated, and then each feature vector x_m in matrix A^k is linearly multiplied by the ratio c_k.

$$c_k = \overline{V} ./ v_k \tag{4}$$

$$\overline{A^k} = A^k .* c_k \tag{5}$$

The next steps are performed on the equalized voiceprint spectrums to obtain the equalization feature, which includes filtering by Mel Triangle Filter Bank filters, taking logarithm and DCT.

3. GRU MODEL

RNN has the ability to deal with a time-dependent data sequence with a strong correlation, where it can memorize the previous information and apply it to get the following information. However, since RNN usually uses tanh as its activation function which can cause gradient vanishing, researchers proposed LSTM. LSTM is an effective technology to solve gradient vanishing, and it has a high universality and a great variety of possibilities. As a variant of LSTM, GRU can simplify the calculation process and shorten the training time while retaining the excellent performance of LSTM. The data flow in GRU is shown in Fig. 6.

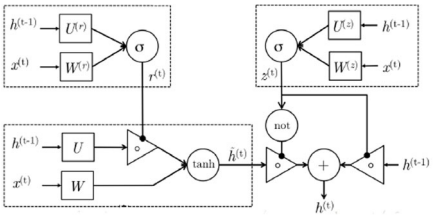


Fig. 6. The data flow in GRU.

The mapping relationship of the data corresponding to Fig. 6 is as follows:

$$z^{(t)} = \sigma(W^{(z)}x^{(t)} + U^{(z)}h^{(t-1)}) \tag{6}$$

$$r^{(t)} = \sigma(W^{(r)}x^{(t)} + U^{(r)}h^{(t-1)}) \tag{7}$$

$$\tilde{h}^{(t)} = \tanh(r^{(t)} \odot Uh^{(t)} + U^{(r)}h^{(t-1)}) \tag{8}$$

$$h^{(t)} = (1 - z^{(t)}) \odot \tilde{h}^{(t)} + z^{(t)} \odot h^{(t-1)} \tag{9}$$

As the deep learning model is applied to a classification task, the classification difficulty of a common method to train a multi-classification model will be aggravated by the increasing number of categories. In this paper, we propose a new method to train the classifier. For a *K*-class language recognition problem, we train *k* parallel binary-classifiers to replace the *K*-class model. For the *k*th binary-classifier, the speech segments of the *k*th language are labeled as positive, and the others are negative. This model determines which language the input segment belongs to by calculating the similarity between the input segment and the *k*th language and the result is the language that the maximum similarity corresponds to.

4. EVALUATION AND DISCUSSION

In order to evaluate the performance of the proposed method, experiments are carried

out on the NIST LRE 2007 corpus. It is a language recognition database with 8 types of languages, where all utterances are recorded in single channel with 16-bit streams at 8000Hz sampling rate. Since the dialect is not researched in this paper, experimental evaluation is fixed on 4 types standard languages which include Arabic, Bengali, Russian and Thai. Each type of language consists of 400 minutes training speech, and three kinds of test set including average durations of 3 seconds, 10 seconds and 30 seconds where each class has 80 utterances.

4.1 Experimental Setup

39 dimensional equalization MFCC coefficients (including 13 equalization MFCC, 13Δequalization MFCC and 13ΔΔequalization MFCC coefficients) are used as acoustic features, using 20-msec-long windows shifted by 10 msec.

In addition, the following features are also extracted to compare with the proposed equalization MFCC.

- **MFCC:** traditional MFCC feature without any robust processing.
- **CMS:** mean subtraction over a sliding window of up to 3 seconds [20] on MFCC feature.
- **CMVN:** mean and variance normalization over a sliding window of up to 3 seconds on MFCC feature.
- **RASTA:** the band pass filter ($N=5$, $G=0.1$, $\rho=0.94$) on MFCC feature.

The following systems are constructed to evaluate language recognition.

- ***I*-vector (Baseline system):**
  The experiments operate on 56-dimensional SDC (optimal configuration 7-1-3-7) which are calculated corresponding to MFCC. We use the training data of NIST LRE 2007 corpus to train the UBM with 1024 Gaussians, the total variability matrix composed of 400 total factors and the *i*-vector of 400 dimensionality by Kaldi Identity Toolbox [http://www.kaldi-asr.org/doc/]. Softmax regression is then used as classification.
- **GRU-based classification + equalization MFCC (Proposed method):**
  In GRU-based system, all models are GRUs with 512 units and the input is 10-second equalization MFCC acoustic feature (the dimension is 1001×39). The learning rate is set with 0.001, the batch size is set with 32 and tanh is used as the activation function.
- **VGG16 based classification + equalization MFCC:**
  In VGG16 [21] system, the specific structure of the convolution network consists of a large number of 3*3 convolution kernels (3*3*64, 3*3*64, max pooling, 3*3*128, 3*3*128, max pooling, 3*3*256, 3*3*256, 3*3*256, max pooling, 3*3*512, 3*3*512, 3*3*512, max pooling, 1024 units, 1024 units). The settings of other parameters are the same with GRU.
- **LSTM based classification + equalization MFCC:**
  All models are left-to-right unidirectional LSTM with 512 units and the settings of other parameters are the same with GRU.
- **Bidirectional LSTM based classification + equalization MFCC:**
  All models are bidirectional LSTM with 512 units and the settings of other parameters are the same with GRU.

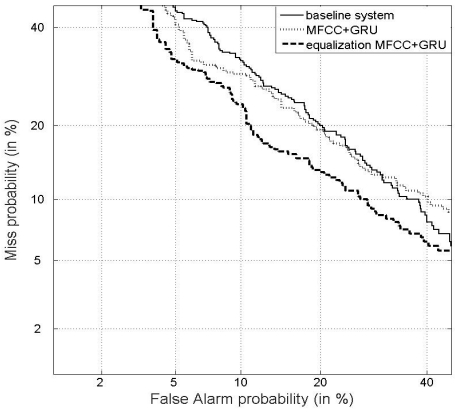


Fig. 7. Performance of different methods on 3s test data in NIST LRE 2007 corpus.

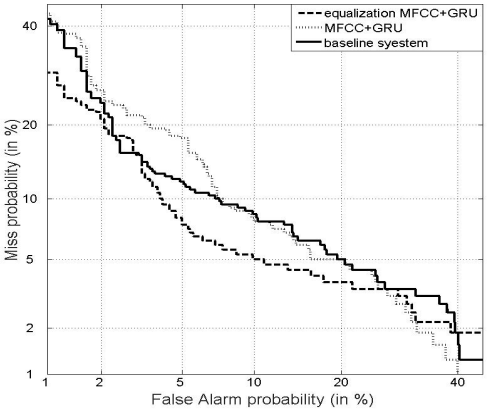


Fig. 8. Performance of different methods on 10s test data in NIST LRE 2007 corpus.

4.2 Results and Discussions

Fig. 7 shows the performance of the proposed method compared with the typical *i*-vector (the baseline system) and MFCC feature plus GRU model (MFCC+GRU) on 3-second test data. The solid line is used to describe the detection error tradeoffs (DETAs) of the baseline system, the dotted line represents the DETs of MFCC+GRU and the dashed line denotes the DETs of the proposed method. It can be seen that the proposed method (EER with 14.74%) significantly outperforms the baseline system (*i*-vector EER with 19.79%) and MFCC+GRU (EER with 18.80%).

Fig. 8 shows the performance of the proposed method compared with the typical *i*-vector and MFCC+GRU on 10-second test data. The solid line is used to describe detection error tradeoffs (DETAs) of the baseline system. Moreover, the dotted line represents the DETs of MFCC+GRU and the dashed line denotes the DETs of the proposed method. It shows that the proposed method (EER with 6.09%) significantly outperforms the baseline system (*i*-vector EER with 8.18%) and MFCC +GRU (EER with 8.44%).

Fig. 9 shows the performance of the proposed method compared with the typical *i*-

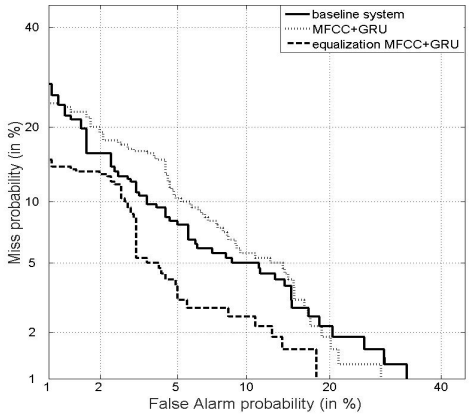


Fig. 9. Performance of different methods on 30s test data in NIST LRE 2007 corpus.

vector and MFCC+GRU on 30s test data. The solid line is used to describe detection error tradeoffs (DETAs) of the baseline system. Furthermore, the dotted line represents the DETs of MFCC+GRU and the dashed line denotes the DETs of the proposed method. It shows that the proposed method (EER with 4.06%) significantly outperforms the baseline system (*i*-vector EER with 6.04%) and MFCC+GRU (EER with 7.50%).

In order to show the experimental performance clearly, Table 1 is used to summarize the results in Figs. 7-9.

In Table 1, the data illustrates that the performance can be better as the time increases. In the same test conditions, Equalization MFCC+GRU is the best compared with the baseline system and MFCC+GRU.

Table 1. Performance of different methods on 3s, 10s and 30s test data in NIST LRE 2007 corpus.

Time	Baseline system	MFCC+GRU	Equalization MFCC+GRU
3s	19.79%	18.80%	14.74%
10s	8.18%	8.44%	6.09%
30s	6.04%	7.50%	4.06%

Table 2. Performance of various classifiers using the equalization MFCC feature in NIST LRE 2007 corpus.

Time	GRU	VGG16	LSTM	Bid-LSTM
3s	14.74%	17.80%	16.35%	17.55%
10s	6.09%	7.78%	9.48%	11.46%
30s	4.06%	8.16%	7.55%	8.33%

In order to show the stability of proposed feature, the equalization MFCC feature pairs with various classifiers such as GRU, VGG16, LSTM and Bid-LSTM in Table 2.

Table 2 illustrates the reason that we use the GRU network as the classifier rather than others. Without too much effort on tuning these non-optimal models, we can achieve a better performance using great parameters.

In order to assess the robustness of the equalization features, we add babble and factory noises to the NIST LRE 2007 test set, with SNR of 0 dB, 5 dB and 10 dB respectively. Fig. 10 shows the test results on three-second test data in NIST LRE 2007 corpus, and lines are used to describe the EER of the baseline system and the new feature system. Under the babble noisy environment (0dB), the result indicates that the EER of the baseline system is 42.55% and that of the new system is 36.46%. Under the babble noisy environment (5dB), the result indicates that the EER of the baseline system is 38.13% and that of the new system is 31.35%. Under the babble noisy environment (10dB), the result indicates that the EER of the baseline system is 33.75% and that of the new system is 25.99%.

Fig. 11 shows the test results on ten-second test data in NIST LRE 2007 corpus, and lines are used to describe the EER of the baseline system and the new feature system. Under the babble noisy environment (0dB), the result indicates that the EER of the baseline system is 39.90% and that of the new system is 34.43%. Under the babble noisy environment (5dB), the result indicates that the EER of the baseline system is 35.21% and that of

the new system is 29.48%. Under the babble noisy environment (10dB), the result indicates that the EER of the baseline system is 31.88% and that of the new system is 23.28%.

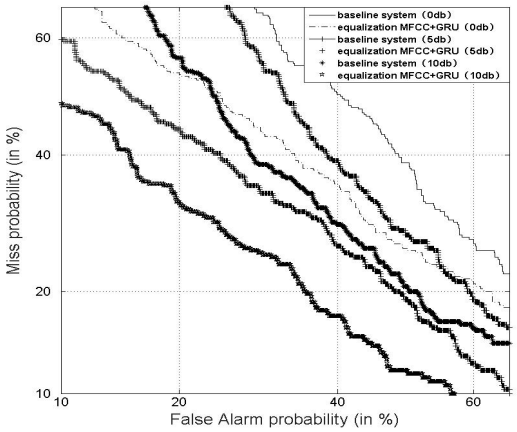


Fig. 10. Performance of different methods under 3 different SNR levels of babble noise conditions on 3s test data in NIST LRE 2007 corpus.

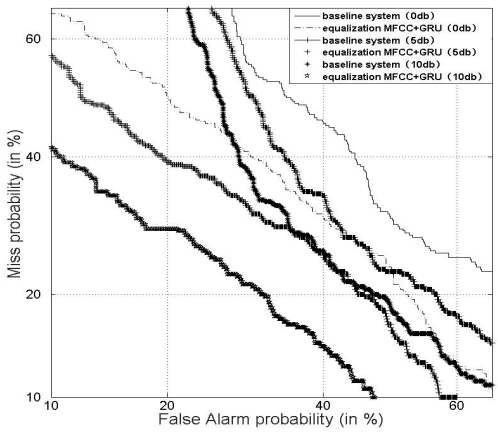


Fig. 11. Performance of different methods under 3 different SNR levels of babble noise conditions on 10s test data in NIST LRE 2007 corpus.

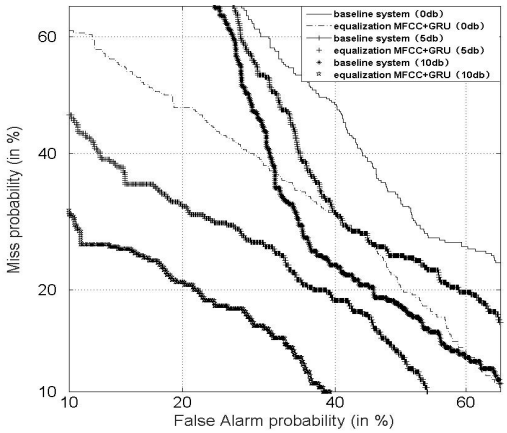


Fig. 12. Performance of different methods under 3 different SNR levels of babble noise conditions on 30s test data in NIST LRE 2007 corpus.

Fig. 12 shows the test results on thirty-second test data in NIST LRE 2007 corpus, and lines are used to describe the EER of the baseline system and the new feature system. Under the babble noisy environment (0dB), the result indicates that the EER of the baseline system is 40.57% and that of the new system is 33.44%. Under the babble noisy environment (5dB), the result indicates that the EER of the baseline system is 34.95% and that of the new system is 24.69%. Under the babble noisy environment (10dB), the result indicates that the EER of the baseline system is 30.73% and that of the new system is 18.39%. In

the noisy conditions, it can be seen that the new feature system is still better than the baseline system.

In order to clearly show the experimental performance under 3 different SNR levels of babble noise conditions on 3s, 10s and 30s test data, Table 3 is used to summarize the results in Figs. 10-12.

Table 3. Performance of different methods under 3 different SNR levels of babble noise conditions on 3s, 10s and 30s test data in NIST LRE 2007 corpus.

Time	SNR level	Baseline system	Equalization MFCC+GRU
3s	0dB	42.55%	36.46%
	5dB	38.13%	31.35%
	10dB	33.75%	25.99%
10s	0dB	39.90%	34.43%
	5dB	35.21%	29.48%
	10dB	31.88%	23.28%
30s	0dB	40.57%	33.44%
	5dB	34.95%	24.69%
	10dB	30.73%	18.39%

In Table 3, the data illustrates that the performance can be better as the higher SNR levels in the same time of the test data. In the same SNR levels, the performance can be better as the time increases in most cases. In the same test condition, Equalization MFCC +GRU is better than the baseline system.

On 3s test data in NIST LRE 2007 corpus, under the factory noisy environment (0dB), the results are shown in Fig. 13, where the lines are used to describe the EER of the baseline system (44.69%) and the new feature system (39.22%) respectively. Under the factory noisy environment (5dB), the results are shown that the EER of the baseline system is 40.16% and that of the new system is 36.2%. Under the factory noisy environment (10dB), the results are shown that the EER of the baseline system is 36.82% and that of the new system is 32.29%.

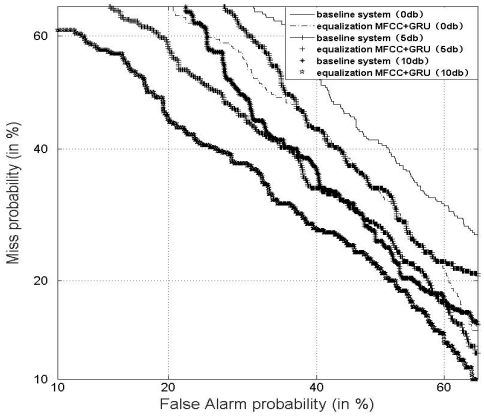


Fig. 13. Performance of different methods under 3 different SNR levels of factory noise conditions on 3s test data in NIST LRE 2007 corpus.

On 10s test data in NIST LRE 2007 corpus, under the factory noisy environment (0dB), the results are shown in Fig. 14, where the lines are used to describe the EER of the baseline system (43.3%) and the new feature system (38.59%) respectively. Under the factory noisy environment (5dB), the results are shown that the EER of the baseline system is 39.27% and that of the new system is 35.78%. Under the factory noisy environment (10dB), the results are shown that the EER of the baseline system is 35.16% and that of the new system is 31.09%.

On 30s test data in NIST LRE 2007 corpus, under the factory noisy environment (0 dB), the results are shown in Fig. 15, where the lines are used to describe the EER of the baseline system (42.24%) and the new feature system (38.96%) respectively. Under the factory noisy environment (5dB), the results are shown that the EER of the baseline system is 38.85% and that of the new system is 35.83%. Under the factory noisy environment (10dB), the results are shown that the EER of the baseline system is 34.84% and that of the new system is 29.64%. In the noisy conditions, it can be seen that the new feature system is still better than the baseline system.

In order to clearly show the experimental performance under 3 different SNR levels of factory noise conditions on 3s, 10s and 30s test data, Table 4 is used to summarize the results in Figs. 13-15.

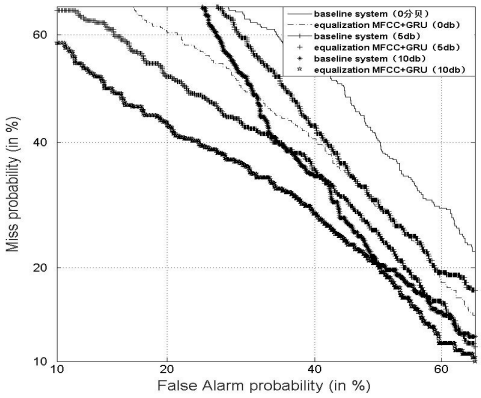


Fig. 14. Performance of different methods under 3 different SNR levels of factory noise conditions on 10s test data in NIST LRE 2007 corpus.

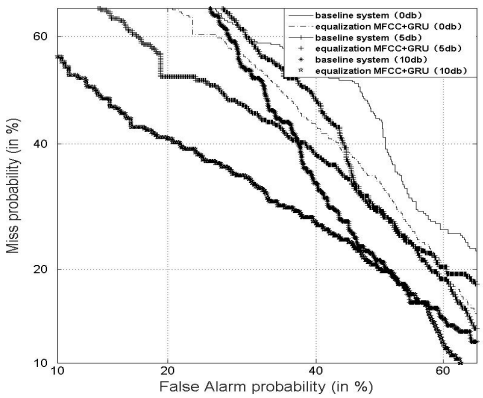


Fig. 15. Performance of different methods under 3 different SNR levels of factory noise conditions on 30s test data in NIST LRE 2007 corpus.

Table 4. Performance of different methods under 3 different SNR levels of factory noise conditions on 3s, 10s and 30s test data in NIST LRE 2007 corpus.

Time	SNR level	Baseline system	Equalization MFCC+GRU
3s	0dB	44.69%	39.22%
	5dB	40.16%	36.20%
	10dB	36.82%	32.29%
10s	0dB	43.30%	38.59%
	5dB	39.27%	35.78%
	10dB	35.16%	31.09%
30s	0dB	42.24%	38.96%
	5dB	38.85%	35.83%
	10dB	34.84%	29.64%

In Table 4, the data illustrates that the performance can be better as the higher SNR levels in the same time of the test data. In the same SNR levels, the performance can be better as the time increases in most cases. In the same test condition, Equalization MFCC+ GRU is better than baseline system.

To further demonstrate the superior robustness of the equalization MFCC feature, we also compared it with other robustness features, such as CMS, CMVN and RASTA, using GRU as the classifier in babble and factory noise conditions.

Tables 5 and 6 illustrate that, among several noise condition set in the experiments, our equalization MFCC feature realizes the best EER in most cases which means it has better robustness.

Table 5. Performance of different methods under 3 different SNR levels of babble noise conditions on 3s, 10s and 30s test data using GRU as classifier.

Time	SNR level	MFCC	Equalization MFCC	CMS	CMVN	RASTA
3s	0db	36.77%	**36.46%**	41.67%	43.80%	46.56%
	5db	**29.58%**	31.35%	38.18%	40.73%	43.28%
	10db	**24.17%**	25.99%	34.27%	38.39%	38.54%
10s	0db	35.99%	**34.43%**	39.84%	45.83%	48.44%
	5db	29.69%	**29.48%**	32.60%	42.92%	46.41%
	10db	23.91%	**23.28%**	26.09%	40.47%	41.25%
30s	0db	38.96%	**33.44%**	37.66%	47.92%	49.27%
	5db	29.53%	**24.69%**	28.49%	46.04%	48.75%
	10db	23.28%	**18.39%**	19.32%	43.70%	44.11%

Table 6. Performance of different methods under 3 different SNR levels of factory noise conditions on 3s, 10s and 30s test data using GRU as the classifier.

Time	SNR level	MFCC	Equalization MFCC	CMS	CMVN	RASTA
3s	0db	43.39%	**39.22%**	43.33%	46.98%	49.27%
	5db	39.38%	**36.20%**	40.63%	44.64%	48.75%
	10db	35.31%	**32.29%**	37.34%	43.59%	44.53%
10s	0db	45.63%	**38.59%**	46.04%	48.96%	49.84%
	5db	40.99%	**35.78%**	38.23%	48.18%	49.48%
	10db	34.53%	**31.09%**	31.98%	46.56%	46.35%
30s	0db	45.63%	**38.96%**	47.92%	50.00%	50.00%
	5db	40.83%	**35.83%**	38.80%	49.69%	49.90%
	10db	34.69%	29.64%	**25.57%**	49.06%	48.28%

Furthermore, the performance of the language recognition is evaluated based on the number of hidden nodes on NIST LRE 2007 corpus. In Fig. 16, the experiments show clearly that different numbers of hidden nodes lead to different results. The accuracy tends to increase first, and gets the best performance at the number of 512 hidden nodes, followed by a decrease as the number of hidden nodes increases. Thus, more numbers of hidden nodes don't mean a better performance.

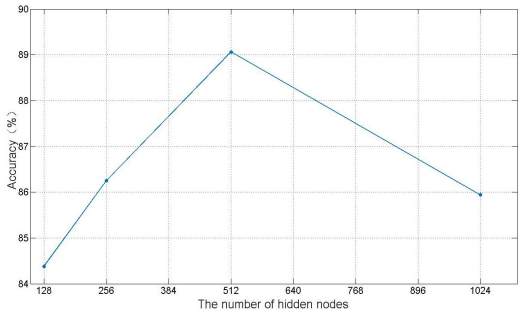


Fig. 16. Comparison of accuracy about the number of different hidden nodes in NIST LRE 2007 corpus.

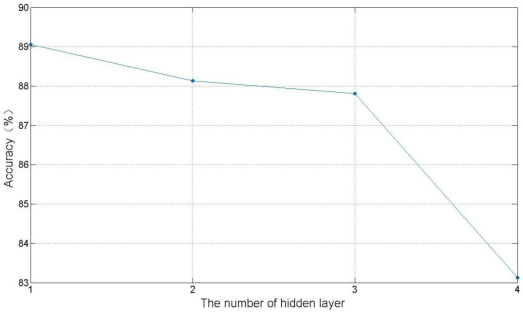


Fig. 17. Comparison of accuracy about the number of different hidden layers in NIST LRE 2007 corpus.

Fig. 17 fixed on 512 hidden nodes when the performance of language recognition evaluated based on hidden nodes shows the best result, and the performance is evaluated based on the number of hidden layers. Several results with the number of different hidden layers are tested on the test set. The highest accuracy of language recognition exhibits at single hidden layer. It may be explained that more numbers of hidden layers don't mean a better performance.

5. CONCLUSIONS

In this paper, we propose the equalization features. The distribution of all the spectrum mean vectors calculated from segments are locally different but have the same trend in the overall distribution, so we apply equalization to these voiceprint spectrums in which the components on every frequency can be expressed clearly. Furthermore, equalization performance can also suppress the noise effect to a certain extent and achieve better robustness compared with other algorithms such as CMS, CMVN and RASTA. In model training, we apply GRU to language recognition. GRU neural network shows great advantages to process time-sequential information of speech compared with other popular neural networks. There are multiple reasons including a simpler structure, faster training speed, and without gradient vanishing. The experimental results prove that the proposed method is able to achieve the better performances than the baseline system in the ideal and noise environments.

ACKNOWLEDGMENTS

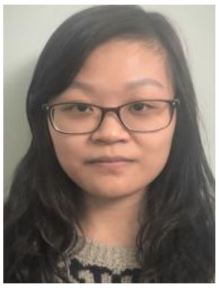
This research is supported by the National key research and development plan (No. 2017YFB0803001), National Science Foundation of China (No. 61571144) and the Fundamental Research Funds for the Central Universities (No. HIT.NSRIF.2019082).

REFERENCES

1. T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervector," *Speech Communication*, Vol. 52, 2010, pp. 12-40.
2. H. Veisi and H. Sameti, "The integration of principal component analysis and cepstral mean subtraction in parallel model combination for robust speech recognition," *Digital Signal Processing*, Vol. 21, 2011, pp. 36-53.
3. B. Bharathi, S. Kavitha, and K. M. Priya, "Speaker verification in a noisy environment by enhancing the speech signal using various approaches of spectral subtraction," in *Proceedings of the 15th Annual International Conference on Intelligent Systems and Control*, 2016, pp. 1-5.
4. V. Joshi, N. V. Prasad, and S. Umesh, "Modified mean and variance normalization: transforming to utterance-specific estimates," *Circuits Systems & Signal Processing*, Vol. 35, 2016, pp. 1593-1609.
5. N. V. Prasad and S. Umesh, "Improved cepstral mean and variance normalization using Bayesian framework," in *Proceedings of Automatic Speech Recognition and Understanding*, 2013, pp. 156-161.
6. L. Singh and G. Chetty, "A comparative study of recognition of speech using improved MFCC algorithms and Rasta filters," *Communications in Computer & Information Science*, Vol. 285, 2012, pp. 304-314.
7. M. A. Islam, W. A. Jassim, and N. S. Cheok, "A robust speaker identification system using the responses from a model of the auditory periphery," *Plos One*, Vol. 11, 2016, pp. e0158520.
8. T. Virtanen, R. Singh, and B. Raj, "Techniques for noise robustness in automatic speech recognition," *West Sussex*, Wiley, UK, 2012.
9. T. N. Trong, V. Hautamäki, and A. L. Kong, "Deep language: a comprehensive deep learning approach to end-to-end language recognition," *Odyssey*, 2016, pp. 109-116.
10. L. Deng, "Deep learning: from speech recognition to language and multimodal processing," *Apsipa Transactions on Signal & Information Processing*, Vol. 5, 2016, pp. 1-15.
11. F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Processing Letters*, Vol. 22, 2015, pp. 1671-1675.
12. X. Chen, X. Liu, and M. J. F. Gales, "Recurrent neural network language model training with noise contrastive estimation for speech recognition," in *Proceedings of the 40th Annual International Conference on Acoustics*, *Speech and Signal Processing*, 2015, pp. 5411-5415.
13. H. Sak, A. W. Senior, and K. Rao, "Fast and accurate recurrent neural network acoustic models for speech recognition," in *Proceedings of the 16th Annual International*

Conference on Speech Communication Association, 2015, pp. 1468-1472.
14. T. H. Kim, "Training method and speaker verification measures for recurrent neural network based speaker verification system," *The Journal of the Korean Institute of Communication Sciences*, Vol. 34, 2009, pp. 257-267.
15. A. Graves, "Long short-term memory," *Supervised Sequence Labelling with Recurrent Neural Networks*, Springer Berlin Heidelberg, 2012, pp. 1735-1780.
16. R. Zazo, A. Lozano-Diez, and J. Gonzalez-Dominguez, "Language identification in short utterances using long short-term memory (LSTM) recurrent neural networks," *Plos One*, Vol. 11, 2016, pp. e0146917.
17. R. Zazo, P. S. Nidadavolu, N. Chen, *et al.*, "Age estimation in short speech utterances based on LSTM recurrent neural networks," *IEEE Access*, Vol. 6, 2018, pp. 22524-22530.
18. K. Berninger, J. Hoppe, and B. Milde, "Classification of speaker intoxication using a bidirectional recurrent neural network," *Text*, *Speech*, *and Dialogue*, Springer, Berlin, 2016, pp. 435-442.
19. J. Kang, W. Q. Zhang, and J. Liu, "Gated recurrent units based hybrid acoustic models for robust speech recognition," in *Proceedings of the 9th Annual International Symposium on Chinese Spoken Language Processing*, 2017, pp. 1-5.
20. Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker 314 embeddings for text-independent speaker verification," in *Proceedings of the 19th Annual International Conference on Speech Communication Association*, 2018, pp. 3573-3577.
21. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2015, pp. 1-13.

Wen-Jie Song (宋文杰) received her B.S. degree in Computer Science and Technology from Harbin Institute of Technology, China, in 2018. In the same year, she was recommended to Harbin Institute of Technology for a master's degree. She is currently a postgraduate student in the Speech Processing Lab, engaged in the research of speech processing.



Chen Chen (陈晨) received her M.E. degree in Computer Science and Technology from Harbin Institute of Technology, China. She is currently a Ph.D. candidate of Harbin Institute of Technology. Her research interests include speaker recognition and language recognition.

Tian-Yang Sun (孙天扬) was admitted into Harbin Institute of Technology in 2015. She is currently a student of Computer Science and Technology, School of Computer and Software. In 2017, she was admitted into the 2+2 Dual degree Program to continue her study in the University of Birmingham, United Kingdom.

Wei Wang (王伟) received her Ph.D. degree in Computer Science and Technology from Harbin Institute of Technology in 2016. She is currently a Lecturer of Computer Science and Technology, Harbin Institute of Technology. Her research interests include speech signal processing and audio information processing.