Journal of Computational Information Systems 11: 9 (2015) 3087–3095
Available at http://www.Jofcis.com

Speaker Identification using Spectrogram and Learning Vector Quantization *

Penghua LI*, Shunxing ZHANG, Huizong FENG, Yuanyuan LI

Automotive Electronics Engineering Research Center, College of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Abstract

A text-independent, closed-set speaker identification method is proposed in this paper. The method uses the textures in spectrogram as the features of speech signal and the learning vector quantization (LVQ) neural network as the feature classifier. The input speech signal is pre-emphasized, windowed, and FFT analyzed, resulting a spectrogram. To reduce the impact of external noise, the textures in spectrogram are enhanced by Gabor filter. The enhanced spectrogram are processed by local binary patterns (LBP) operator for getting a low-dimensional feature vector, which alleviates the computational burden of the classifier. Then the LBP vectors are fed to LVQ neural network for the classification of speaker identification. The numerical experiments are carried out to verify the theoretical results and clearly show that our identification method has good recognition ability in term of accuracy.

Keywords: Speaker Identification; Spectrogram; Learning Vector Quantization; Gabor Transform

1 Introduction

Speaker recognition, a hard and practical work, has two parts: verification and identification. The verification handles with confirming the identity claim of a speaker, while the identification conducts which one of the given voices to be best matched with the input voice sample [1]. The speaker identification, with the purpose preventing a single person to use multiple identities [2], is further classified into "open-set" and "closed-set" tasks. If the target speaker is assumed to be one of the registered speakers, the recognition task is a "closed-set" problem. If not, the target speaker is none of the registered speakers, the task belongs to "open-set".

Actually, the process of speaker identification is composed of the feature extraction in front-end and the mode classification in back-end. So far, several kinds of well-established methods have been proposed for the feature extraction, such as the mel frequency cepstral coefficients (MFCCs),

*Project jointly supported by the National Natural Science Foundation (61403053), the Youth Science and Technology Innovation Talents Project of Chongqing (cstc2013kjrc-qnrc40005, CSTC2013kjrc-tdjs40010, cstc2012jjA60002), the Science and Technology Project of Chongqing Municipal Education Commission (KJ1400404).

*Corresponding author.
Email address: lipenghua88@163.com (Penghua LI).

1553–9105 / Copyright © 2015 Binary Information Press
DOI: 10.12733/jcis13619
May 1, 2015

the linear prediction cepstrum coefficients (LPCCs) [3], and so on. Especially, the spectrogram is an effective feature extraction technique and has been given more and more attention by large number of researchers. In [4], the author presents an approach to text dependent speaker identification while an individual utters a given word through the optimally segmented spectrograms. A simple approach to text dependent speaker identification using spectrograms and row mean is presented in [5]. The authors in [6], presents a feature extraction technique for speaker recognition using Radon transform and discrete cosine transform.

In fact, the features exhibited in speech spectrogram are similar to the texture features in the view of image processing. Therefore, a lot of texture processing methods can be used for the analysis of speech spectrogram. Among these methods, the local binary pattern (LBP) method [7] has been widely used to process the texture features because of its excellent property of gray scale invariant. However, the texture in spectrogram is not always clear since the speech signal is contaminated by the environment noise during the sampling process. Thus it is necessary to increase the recognizable degree of these contaminated textures via some image enhancement technique, such as the Gabor transform [8]. On the other hand, the design of back-end classifier have a significant impact on the recognition accuracy. There are a large number of techniques, i.e., the gaussian mixture model (GMM), the hidden Markov model (HMM), to mention but a few, are proposed for speaker identification. Besides these techniques, the neural network (NN) has been playing an increasing role in speech processing, especially for text-independent speaker identification. As a supervised learning technique that can classify input vectors based on vector quantization, the learning vector quantization (LVQ) neural network developed by Kohonen [9], being very similar to the Kohonen SOM, is widely used in pattern recognition.

In this paper, we aim to investigate a text-independent, closed-set speaker identification method using spectrogram and LVQ network. The input speech signals are pre-emphasized, windowed, and FFT analyzed, resulting in a spectrogram. To reduce the impact of external noise, the textures in spectrogram are enhanced by Gabor filter. The enhanced spectrogram are processed by local binary patterns (LBP) operator for getting a low-dimensional feature vector, which alleviates the computational burden of the classifier. Then the LBP vectors are fed to LVQ neural network for the classification of speaker identification.

2 Proposed Speaker Identification Approach

The proposed speaker identification approach, being shown in Fig. 1, is composed of three parts: the preprocessing module, the feature extraction module and the classification module.

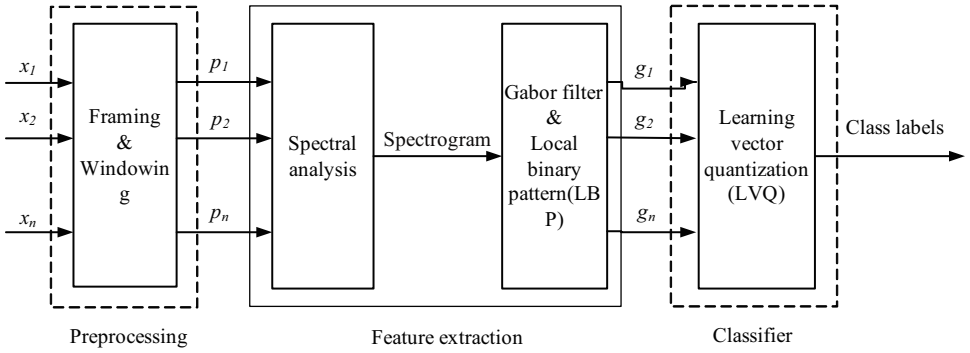

Fig. 1: The proposed speaker identification approach

The utterances are pre-emphasized and windowed in the preprocessing module. In the feature extraction module, the preprocessed utterances transformed by short-time FFT to obtain the spectrogram, then fed to the Gabor filter for texture enhancement, and computed by LBP algorithm to get feature vectors. In the classification module, these lower-dimensional features are submitted to the LVQ network for the training and testing of the speaker identification.

2.1 Spectrogram

Spectrogram is usually created via the calculation of time signal using the FFT, which is approximated as a filter bank that results from a series of band-pass filters. Specifically, the time and the frequency attributes of a speech signal are indicated separately by the cross and the vertical axes in the corresponding spectrogram. In addition, the gray value of each pixel in the two-dimensional spectrogram reflects the energy of the speech signal at corresponding time point and frequency domain.

The energy power spectrum is given as:

$$P_x(n,\omega) = \frac{1}{2N+1}\left|X(n,\omega)\right|^2 \tag{1}$$

$$X(n,\omega) = \sum_{k=-\infty}^{\infty} x[k]\omega[n-k]\, e^{-j\omega k} \tag{2}$$

where $\omega[n]$ denotes a window function with length of $2N+1$, and $X(n,\omega)$ represents the transformed value of a frame signal, with a central point n in time domain, which has performed the FFT at ω. In practice, it is not necessary to calculate the energy value of a speech signal at each possible frequency point or time point. Therefore, it is enough to calculate $2N+1$ points in the frequency domain, or calculate N points in the time domain.

2.2 Gabor transform

The Gabor transform, a variant of the Fourier transform, provides a scope judgment window for texture image enhancement. The Gabor function can capture a considerable number of texture information, with excellent spatial and the joint frequency of resolution. Specifically, the enhanced features of different scales and directions in the frequency domain can be extracted by the Gabor function.

A two-dimensional Gabor function is described as following:

$$g_{uv}(x,y) = \frac{k^2}{\sigma^2}\exp\left[-\frac{k^2(x^2+y^2)}{2\sigma^2}\right] \bullet \left\{\exp\left[ik \bullet \begin{pmatrix} x \\ y \end{pmatrix}\right] - \exp\left(-\frac{\sigma^2}{2}\right)\right\} \tag{3}$$

where

$$k = \begin{pmatrix} k_x \\ k_y \end{pmatrix} = \begin{pmatrix} k_v \cos\varphi_u \\ k_v \sin\varphi_u \end{pmatrix} \tag{4}$$

$$k_v = 2^{-\frac{v+2}{2}}\pi \tag{5}$$

$$\varphi_u = u\frac{\pi}{k}. \tag{6}$$

Here, the values of "v" and "u" represent the wavelength and the kernel function direction of the Gabor wavelet, respectively. The total quantity of direction is determined by the value of "k". The parameter of σ/k determines the size of the Gaussian window, and $\sigma = \sqrt{2}\pi$.

2.3 Local binary patterns operator

The local binary patterns (LBP) operator, extracting the local gray-scale information of the image, is a powerful texture measure with a low-computational complexity. This operator is given by

$$LBP_{P,R}(x,y) = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \tag{7}$$

where

$$s(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases} \tag{8}$$

During the LBP operation, a threshold of each neighbor, g_p $(p = 0, 1..., P-1)$ is obtained by the application of the value of current pixel, g_c, which is represented by a binary number. Concatenating these binary numbers, then converting the sequence into the decimal number, a local binary pattern is obtained. The choice of any radius, R, and number of pixels in the neighborhood, P, to form an operator is allowed by using circular neighborhoods and linearly interpolating the pixel values. In addition, the spot, flat area, edge and corner of the texture image can be represented by a subset of 2^P binary patterns. The uniform LBP operator is given as following:

$$LBP^{\mu_2}_{P,R}(x,y) = \begin{cases} I(LBP_{P,R}(x,y)) & if \quad U(LBP_{P,R}) \leq 2, I(z) \subseteq [0,(P-1)P+1] \\ (P-1)P+2 & otherwise \end{cases} \tag{9}$$

where

$$U(LBP_{P,R}) = |s(g_{P-1} - g_c) - s(g_0 - g_c)| + \sum_{p-1}^{P} \left|s(g_p - g)_c - s(g_{P-1} - g_c)\right| \tag{10}$$

The superscript μ_2 in Eq. (9) indicates that the definition relates to uniform patterns with a U value of at most 2. If $U(x)$ is smaller than 2, the current pixel will be labeled by an index function, $I(z)$. Otherwise, it will be labeled as $(P-1)P+2$. The index function, $I(z)$, containing $(P-1)P+2$ indices, is used to assign a particular index to each of the uniform patterns.

2.4 Learning vector quantization

Learning vector quantization (LVQ), developed by Kohonen [9], is a supervised learning technique that can classify input vectors based on vector quantization. The LVQ training process proceeds with an input vector being randomly selected (along with the correct class for that vector, thus the

supervised learning) from the "labeled" training set. There can actually exist several reproduction (prototype) vectors per class or category.

Given an input vector x_i to the network, the "output neuron" (i.e.,the class or category) in LVQ is deemed to be a "winner" according to

$$\min_{\forall j} d(x_i, w_j) = \min_{\forall j} ||x_i - w_j||_2^2 \tag{11}$$

We let the set of input vectors be denoted as $\{x_i\}$, for $i = 1, 2, ..., N$, and the network synaptic weight vectors (Voronoi vectors) are denoted as $\{x_j\}$, for $j = 1, 2, ..., m$. We also let C_{w_j} be the class (or category) that is associated with the (weight) Voronoi vector w_j, and C_{x_i} is the class label of the input vector x_i to the network. The weight vector w_j is adjusted in the following manner:

(1) If the class associated with the weight vector and the class label of the input are the same, that is, $C_{w_j} = C_{x_i}$, then

$$w_j(k+1) = w_j(k) + \mu(k)[x_i - w_j(k)], \tag{12}$$

where $0 < \mu(k) < 1$ (the learning rate parameter).

(2) But if $C_{w_j} \neq C_{x_i}$, then

$$w_j(k+1) = w_j(k) + \mu(k)[x_i + w_j(k)], \tag{13}$$

and the other weight vectors are not adapted.

3 Numerical Experiments

In this section, we report the results of speaker identification using the proposed method. The experiments are performed on a speech database collected from 5 (3 males and 2 female) healthy adult Chinese speakers. Each speaker repeated an assigned sentence 10 times, and there are 7 sentences in all, as shown in Table 1.

Table 1: The sentences in the speech recoding experiment

Sentence Order	Chinese Phonetic	English Representation
1	bāng zhù	Help
2	duō méi tí	Multi-media
3	fǒu	No
4	kōng tiáo	Air-condition
5	qú xiāo	Cancel
6	shì	Yes
7	shōu yīn jī	Radio

The aforementioned utterances are recorded by a microphone with a data acquisition system. In the proposed identification system, speech signals are framed by a series of Hamming windows.

The frame size is 30 ms and the overlapped length is 15 ms. The measured maximum frequency is 8 kHz while the sampling rate is set as 16 kHz. These time domain speech signals, taking the 7 sentences said by a female speaker as an example, are plotted in Fig. 2.

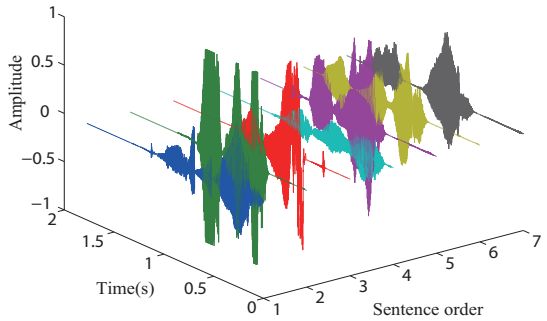

Fig. 2: Speech signals of 7 sentences in the time domain

To obtain the texture feature, all the pre-processed utterances are transformed by FFT, which are described by spectrograms. For instance, a group of spectrograms, representing one sentence spoken by five different persons, are shown in Fig. 3. It is well-known that the speech feature with good separability has the following properties: the features of the same person should be similar, and the features of different persons should be different. However, a large number of similarities exist in the spectrograms shown in Fig. 3(a)-(d), yet many differences exit in the same person plotted in Fig. 3(e)-(f), which is not conducive to distinguish the speaker features. In addition, there are significant vertical stripes displayed in Fig. 3, which represents the interference of the noise. Therefore, it is necessary to enhance the texture in these images and extract further speech features with good separability via another effective techniques.

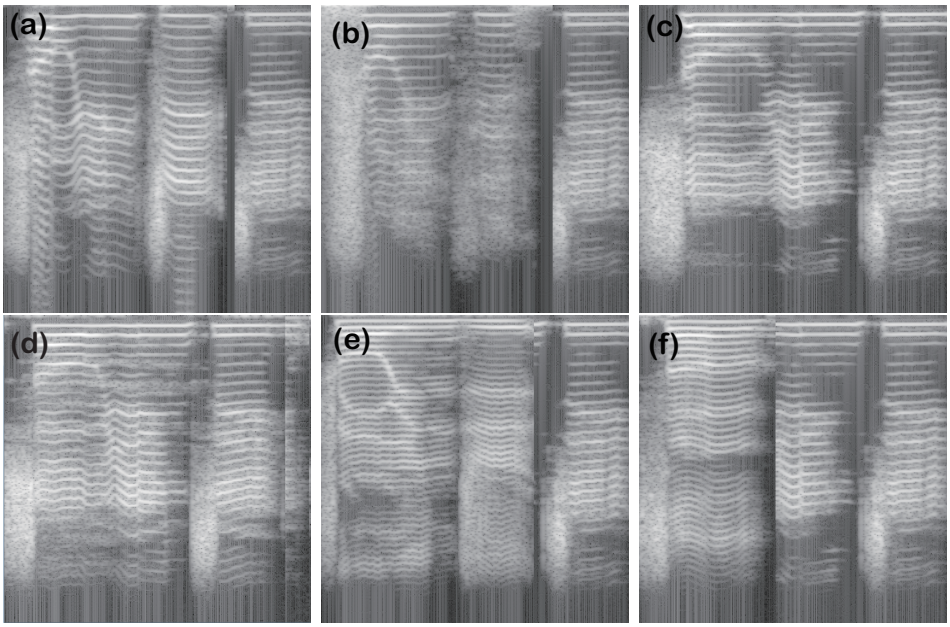

Fig. 3: Spectrograms of the utterance “duō méi tǐ”. (a)-(b) are the spectrograms extracted from two females; (c)-(d) are the spectrograms extracted from two males; (e)-(f)are the spectrograms extracted from a same male

The Gabor filter, an image enhancement technique, is applied to obtain a clearer texture in the spectrogram. Setting the direction parameter σ to be different values, we can get a series of textures images with different degree of recognition. The filtered texture images, taking the texture feature shown in Fig. 3(e) as an example, are shown in Fig. 4.

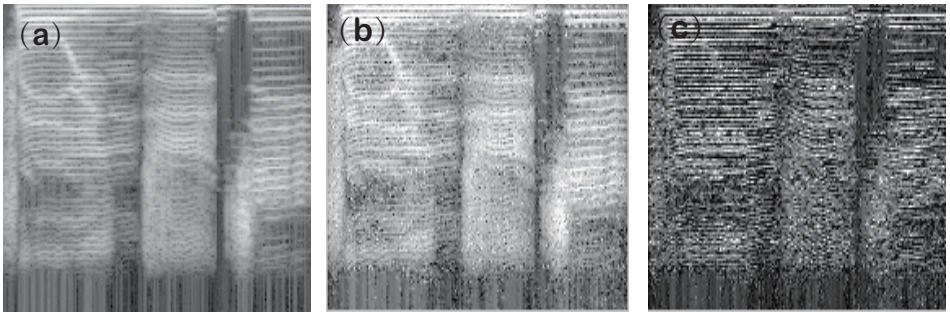

Fig. 4: The comparison between the original spectrogram and the Gabor filtered spectrogram in different direction. (a) the original spectrogram; (b)-(c) are the Gabor filtered spectrogram, where the parameters are set as $\sigma = \pi/6$ and $\sigma = \pi/10$, respectively

It shows intuitively that the textures of Fig. 4(b)-(c) are more stereoscopic than those shown in Fig. 4(a). To further analyze the performance of the spectrogram processed by Gabor filter, the mean, contrast and entropy of these feature images are given in Table 2. For both of the mean and the contrast, as shown in Table 2, there are little difference between the Gabor filtered spectrogram with direction parameter $\sigma = \pi/6$ and the other with $\sigma = \pi/10$. However, the entropy of the filtered spectrogram with $\sigma = \pi/6$ is lower than the spectrogram with $\sigma = \pi/10$. Specifically, the former is 0.1298, but the latter is 0.4327. According to a well-known rule that the information embedded in a spectrogram is larger when the entropy of the spectrogram is larger, we choose the the filtered spectrogram with direction parameter $\sigma = \pi/10$ to be the ultimately selected feature.

Table 2: The details of the spectrograms processed by Gabor filter

Direction Parameters	Mean	Contrast	Entropy
$\sigma = \pi/6$	0.0039	1.0698	0.1298
$\sigma = \pi/10$	0.0040	1.0721	0.4327

After the processing of the Gabor filter, the spectrograms are processed by the LBP operator to obtain the feature vectors, then the feature vectors are fed to LVQ neural network for the speaker identification. It is noting that there are two recognition modes for the training and testing of the network: (1) the corpus take the form of a separate utterance; (2) the corpus take the form of combined utterances. For example, we use the second sentence "duō méi tí" and the combination of the first sentence "bāng zhù" and the fourth sentence "kōng tiáo" to achieve the two recognition modes, respectively. The identification results, described by two confusion matrixes, are given in Fig. 5.

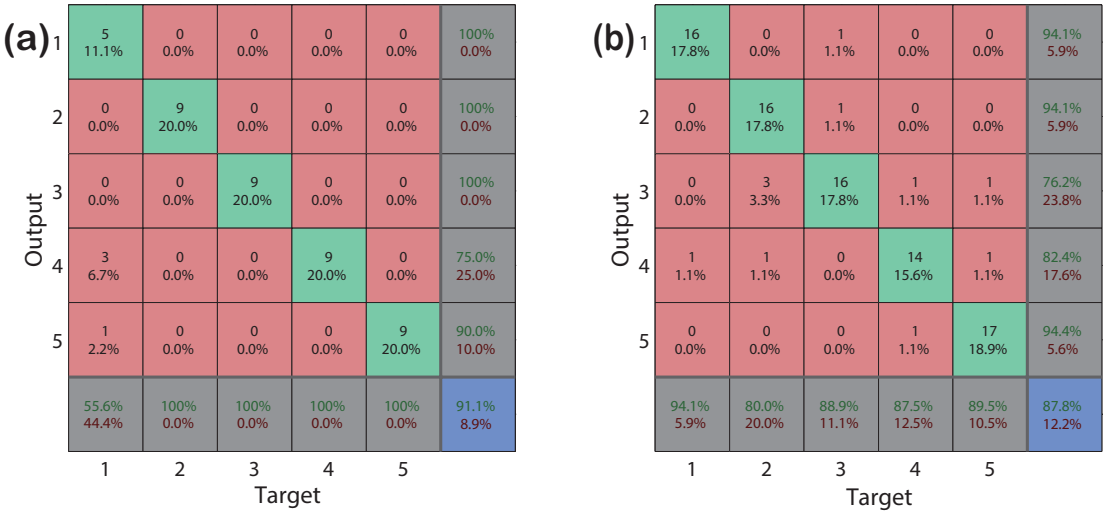

Fig. 5: The confusion matrixes. (a) the result of single utterance, (b) the result of combination of two sentences

To evaluate the performance of the proposed identification method, several similar experiments are carried out using the remaining corpus. The identification results are give in Table 3.

Table 3: The identification results using single and combined utterances.

Order	Sentence	Identification Rate	Combined Sentences	Identification Rate
1	bāng zhù	92.2%	1 & 4	86.7%
2	duō méi tí	91.1%	3 & 4	85.6%
3	fǒu	93.3%	3 & 6	87.8%
4	kōng tiáo	92.2%	2 & 7	83.3%
5	qú xiāo	91.1%	4 & 5	85.6%
6	shì	93.3%	1 & 5	88.9%
7	shōu yīn jī	90.0%	1 & 2	87.8%

As shown in Table 3, the identification rates using single utterance are higher than those using combined utterances. In addition, the recognition rates of the identifiable manner using single sentence can achieve more than 90%, which illustrate adequately that the proposed identification method has a good performance.

4 Conclusion

This paper has focused on the scenario of a systematic approach to perform speaker identification based on spectrograms and LVQ neural network. The spectrograms, being obtained by FFT, are enhanced via Gabor filter to depict the features of speech signals. To reduce the training data and time cost of the LVQ network designed as the feature classifier, the enhanced spectrograms

are processed by LBP operator. Then these LBP vectors are fed to the LVQ neural network for speaker identification. The results indicate that the proposed method has an acceptable recognition rate with high accuracy.

References

[1] Jawarkar, N. P., Holambe, R. S., & Basu, T. K. (2014, February). On the use of classifiers for text-independent speaker identification. In Automation, Control, Energy and Systems (ACES), 2014 First International Conference on (pp. 1-6). IEEE.

[2] Xu, L., & Yang, Z. (2013, August). Speaker identification based on sparse subspace model. In Communications (APCC), 2013 19th Asia-Pacific Conference on (pp. 37-41). IEEE.

[3] Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. Acoustics, Speech and Signal Processing, IEEE Transactions on, 29(2), 254-272.

[4] Dutta, T. (2008, May). Dynamic time warping based approach to text-dependent speaker identification using spectrograms. In Image and Signal Processing, 2008. CISP'08. Congress on (Vol. 2, pp. 354-360). IEEE.

[5] Kekre, H. B., Athawale, A., & Desai, M. (2011, February). Speaker identification using row mean vector of spectrogram. In Proceedings of the International Conference & Workshop on Emerging Trends in Technology (pp. 171-174). ACM.

[6] Ajmera, P. K., Jadhav, D. V., & Holambe, R. S. (2011). Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram. Pattern Recognition, 44(10), 2749-2759.

[7] Carevic, D., & Caelli, T. (1996, August). Adaptive Gabor filters for texture segmentation. In Pattern Recognition, 1996., Proceedings of the 13th International Conference on (Vol. 2, pp. 606-610). IEEE.

[8] Kawady, T. A., Elkalashy, N. I., Ibrahim, A. E., & Taalab, A. M. I. (2014). Arcing fault identification using combined Gabor Transform-neural network for transmission lines. International Journal of Electrical Power & Energy Systems, 61, 248-258.

[9] Kohonen, T. Learning vector quantization for pattern recognition. Techincal Report TKK-F-A601, Helsinki University of Technology, Finland, 1986.