Robust Speaker Verification Based on Max Pooling of Sparse Representation

Wei Wang[1] Jiqing Han[1*] Tieran Zheng[1] Guibin Zheng[1]

[1] School of Computer Science and Technology, Harbin Institute of Technology

Harbin, China

wangwei925229@sohu.com, jqhan@hit.edu.cn, zhengtieran@hit.edu.cn, zhengguibin@hit.edu.cn

Received 4 August 2013; Revised 16 October 2013; Accepted 30 December 2013

Abstract. In the human nervous system, sensory inputs are coded in a sparse manner where only small numbers of neurons are active at a given time, thus the sparse coding is reasonable to be as a plausible model of the auditory cortex. In this paper, we propose a biologically inspired feature extraction method for speaker verification based on sparse coding. When encoding the speech data using sparse coding model, the learned dictionary has the similar characteristics with simple cell receptive fields of auditory neurons and the sparse coding coefficients simulate the response of the auditory cortex neuron. Moreover, every dictionary is learned from every speaker training sample, so that it has more individual information of the speaker and is useful for discriminating different speakers with less dictionary atoms. And based on human auditory masking effect, a neuron which performs a Max Pooling operation on the pooled inputs responds to the strongest one of its inputs and inhibits other weaker inputs. The robustness of the proposed method is better in terms of a strategy to represent natural sounds. The experimental results show that the proposed method outperforms the baseline system on two typical corpuses.

Keywords: speaker verification, sparse representation, robust feature extraction

1 Introduction

Speaker verification is a sort of biometrics technology which aims to determine whether a person's claimed identity is correct or whether the person is an imposter. This technology is widely used in many identity validation fields such as the entrance guard system, telephone banking, and database accessing. At present, the main research focuses on solving the mismatch problem between training and testing samples. Different techniques have been proposed to address this problem, such as the feature extraction technique [1], the model-based transformation technique [2, 3] and score normalization technique [4, 5]. The extraction of effective feature representation plays a crucial role in solving the mismatch problem. As the current principal feature extraction technique, Mel-Frequence Cepstral Coefficient (MFCC) has been successfully used in the speaker verification and achieves the desired performance in the ideal environments, but its performance is deteriorated rapidly in the noisy situations [6]. Therefore, it is necessary to explore a new robust feature extraction technique for the speaker verification in different environments.

It is well known that the human auditory system possesses remarkable abilities to detect, separate, and recognize the speech, the music, and other environmental sounds. Therefore, humans outperform the best machine audition systems by almost any measure, and it is an attractive idea to build a system that emulates speaker recognition in cortex. Psycho-physiological investigations [7, 8] indicate that the acoustic stimulus only need activate a small number of cortical neurons in the brain. Based on this evidence, the sparse coding theory is developed to encode acoustic signals efficiently, so that it can be used to search for the most compact representation of signals in terms of the linear combination of atoms in an overcomplete dictionary [9]. There are two methods to learn the dictionary for the sparse coding including the data model (e.g., wavelets [10], curvelets [11] and Gabor functions [12]) and the data-driven approaches (e.g., k-SVD [13] and online dictionary [14]). When standard atoms are chosen based on the data model, the dictionary atoms do not have any particular semantic meaning. In contrast, the dictionary learned by data-driven approaches has more flexibility to represent natural signals, which have been successfully used in various applications [9, 15-17]. For instance, one exemplar dictionary can be created by composing the training samples of all classes, and the classification is performed by comparing with the norm values or the residual errors of sparse coding coefficients between the unknown speaker and object speaker [15]. However, it has much redundancy by using the training samples as the dictionary. In addition, if the training samples are huge, the computation will be an intractable problem. Therefore, a more compact and/or robust dictionary need be learned from the training samples. Another exemplar dictionary using the GMM mean supervectors of all training speaker utterance can also be applied to the speaker verification by comparing the norm values of sparse coding coefficients between the unknown speaker and object speaker [15,

18]. However, the speaker models are constructed by a large number of mixtures Gaussian models. As a result, such large size dictionaries are not feasible for the greedy search with limited computational resources. On the other hand, when training samples are clustered to obtain parameter estimations, the computation can be reduced, but the valuable information of individual training samples will be lost which affects the performance. Therefore, it is helpful to track the individual information of training samples. In this paper, every dictionary is learned from every training sample, so that it has more individual information of the speaker and is useful for discriminating different speakers with less dictionary atoms. The learned dictionary has the similar characteristics with simple cell receptive fields of auditory neurons and the sparse coding coefficients simulate the response of the auditory cortex neuron when encoding the speech data. Based on the human auditory masking effect, Max Pooling procedure on the sparse coding coefficients is well established. For audio signals, both biological and psychoacoustic evidence suggest that humans have a pooling mechanism within critical bands and loudness pooling (cube root of intensity) across bands [19], so the robustness of the proposed method is high in terms of emulating speaker recognition in the cortex.

2 Speaker Verification Based on Sparse Representation

Since humans outperform the best machine auditory systems by nearly any method, the construction of artificial recognition systems has been an attractive idea that emulates our auditory recognition in cortex. If the artificial system could solve what the auditory system does, the recognition system would be an effective method in understanding object recognition. In human nervous system, sensory information is coded in a sparse manner where only small numbers of neurons are active at a given time so that the sparse coding is effective to be as a believable model of the auditory cortex. At present, the sparse coding has been prevalent in the neural sensory systems. A sparse coding is a high dimensional vector that includes most zeros and a few non-zero entries. However, only a few non-zero entries in the sparse coding provide not only a powerful representation that can capture complex structures in data, but also a computational efficiency. Because the sparse representation has naturally discriminative abilities, it has been exploited in various areas of the pattern recognition such as the face recognition, texture classification and speaker recognition. For instance, the face recognition is based on the sparse representation classification (SRC) with an exemplar dictionary which is created by arranging the training samples of all classes as columns. The test data is represented as the sparse linear combination of the atoms (columns) of the dictionary. The test speaker is assigned to the class which is associated to the atom of the highest non zero coefficient in the sparse vector. Later the similar exemplar dictionary with SRC is created using GMM mean supervectors [15] and the total variability i-vectors [20] for speaker verification task. Different from exemplar dictionaries, the learned dictionaries not only outperform the examples but also are more data-independent [21].

In this paper, the key design for our proposed method based on learned dictionaries has two aspects. One is that every dictionary is learned from every training sample, so that it has more individual information of the speaker and is useful for discriminating different speakers with less dictionary atoms. Secondly, using the auditory models to emulate speaker recognition in the cortex, our system follows the biologically inspired features which use human masking effect in auditory cortex because the learned dictionary has the similar characteristics with simple cell receptive fields of auditory neurons. The sparse coding coefficients simulate the response of the auditory cortex neuron to encode the speech data and the Max Pooling operation on the sparse coding coefficients is established based on human masking effect in auditory cortex. The human masking effect means that the weak signal is inaudible in the vicinity of a strong signal where a neuron which performs a Max Pooling operation on the pooled inputs corresponds to the strongest one of its inputs, and inhibits other weaker inputs. Therefore, our proposed approach is strongly robust to noises based on the auditory models.

2.1 Auditory Masking Effect

Auditory masking is an interesting psychoacoustic phenomenon of the human hearing system. In the presence of a strong sound, many weaker sounds get masked [22]. In Fig. 1, the auditory masking effect phenomenon is described, where x axis represents the frequency range of the human ear and y axis represents the sound pressure level. Normal human ears have a dynamic frequency range from about 20 to 20000 Hz. In a quiet environment, the human ear can perceive the hearing threshold which is the minimum sound pressure to different frequencies sound. When the masker exits at about 500 Hz, the hearing threshold is changed and three sounds at about 380 Hz, 680 Hz and 720 Hz are masked. i.e., these sounds will not be audible. Because the auditory masking effect exits, three masked sounds are lower than the masking threshold and higher than the hearing threshold, which are not perceived.

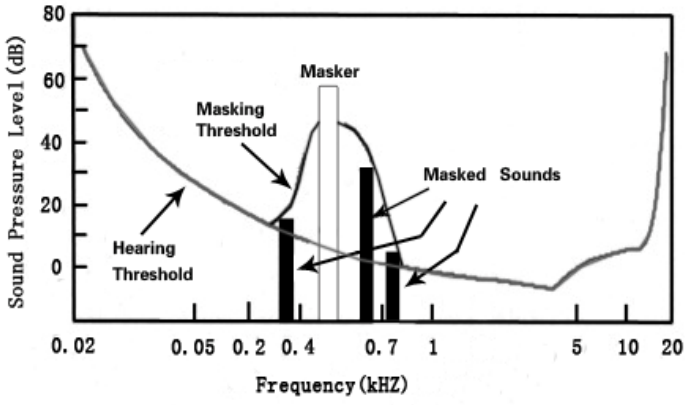

Fig. 1. Auditory masking effect

Masking is used to investigate the auditory system's ability to separate the components of a complex sound. For example, if two sounds of two different frequencies exit at the same time, they can be heard separately rather than a combination sound, which is known as frequency selectivity. Because of the cochlea filtering, a complex sound is split into different frequencies components which cause a spike at a specific place on the basilar membrane of the cochlea. These components are coded independently on the auditory nerve which transmits sound information to the brain.

2.2 Sparse Coding of Speech

We define a speech frame feature sequence $X=[x_1,...,x_N]\in R^{K\times N}$ with $x_n=[\mathrm{x}_{1n},...,\mathrm{x}_{Kn}]^T$ where there are $n=1,...,N$, the signal dimension K, the number N of the speech frames and the transposition T. $D=[d_1,...,d_M]\in R^{K\times M}$ ($K<M$) consisting of M atoms is the base (dictionary) with $d_m=[\mathrm{d}_{1m},...,\mathrm{d}_{Km}]^T$ in which there are $m=1,...,M$, $\forall m$ and $d_m\in R^M$ with $\| d_m \|_2=1$ and $\alpha=[\alpha_1,...,\alpha_N]\in R^{M\times N}$ is the sparse coding coefficient.

The sparse coding [23] is an unsupervised model in which the unlabeled input data X are represented the linear combinations of these basis vectors d_m with the additive noise ε and

$$X=\sum_{m=1}^{M}\alpha_m d_m+\varepsilon=D\alpha+\varepsilon \tag{1}$$

Sparse coding is designed to study the production model $p(X\mid D)$, so that the distribution $p(X\mid D)$ is as close as possible to the empirical distribution of the input data $p^*(X)$. The production model $p(X\mid D)$ is decomposed into

$$p(X\mid D)=\int p(X,\alpha\mid D)d\alpha=\int p(X\mid\alpha,D)p(\alpha)d\alpha \tag{2}$$

If the noise ε follows the Gaussian white noise distributions, there is

$$p(X\mid\alpha,D)=\frac{1}{Z}e^{-\frac{\|X-D\alpha\|_2^2}{2\delta^2}} \tag{3}$$

where δ^2 is the variance of the noise, Z is the normalized constant and $\| X-D\alpha \|_2^2$ is the residue of the linear representation. In order to get the distribution $p(X\mid D)$, it needs to specify the prior distribution $p(\alpha)$. Assuming the sparse coefficients statistically independent, the prior probability can be represented as

$$p(\alpha)=\prod_{m=1}^{M}p(\alpha_m) \tag{4}$$

The specified coefficient distribution is

$$p(\alpha_m) = \frac{1}{Z_\beta} e^{-\beta S(\alpha_m)} \tag{5}$$

where the function $S(\alpha_m)$ determines the shape of the prior distribution, β controls the kurtosis degree of distribution and Z_β is the normalized constant. For instance, if $\beta = 1$ and $S(\alpha_m) = \log(1 + \alpha_m^2)$ are satisfied, $p(\alpha_m)$ follows the Cauchy distribution.

The KL-distance is used to measure the similarity between the production model $p(X \mid D)$ and the empirical distribution of the input data $p^*(X)$, where there is

$$KL(p(X \mid D), p^*(X)) = \int p^*(X) \log \frac{p^*(X)}{p(X \mid D)} dX = \int p^*(X) \log p^*(X) dX - \int p^*(X) \log p(X \mid D) dX \tag{6}$$

Because the empirical distribution $p^*(X)$ is constant and the KL-distance is minimum, this is equivalent to maximizing the log-likelihood of $p(X \mid D)$ in which there is

$$D^* = \arg\max_D < \log p(X \mid D) > \tag{7}$$

Where, $< \log p(X \mid D) >$ denotes expectation over the input data.

However, the integral over α to obtain $p(X \mid D)$ is generally intractable, so we can approximate its integral with the maximum value of $p(X \mid D)$ and obtain an approximate solution

$$D^{*'} = \arg\max_D < \max_\alpha \log p(X \mid D) >= \arg\max_D < \max_\alpha \log p(X \mid \alpha, D) p(\alpha) > \tag{8}$$

Finally, the energy function is defined with

$$E(X, \alpha \mid D) = -\log p(X \mid D, \alpha) p(\alpha) \tag{9}$$

Since maximizing the log-likelihood $p(X \mid D)$ is equivalent to minimizing the energy function, the equation (8) is equivalent to

$$D^{*'} = \arg\min_D < \min_\alpha E(X, \alpha \mid D) > \tag{10}$$

The energy function is further expanded with

$$E(X, \alpha \mid D) = \left\| X - D\alpha \right\|_2^2 + \lambda \sum_{m=1}^{M} S(\alpha_m) \tag{11}$$

where $\lambda = 2\delta^2 \beta$ and the energy function contains two parts. The first part is the reconstruction error, where the base functions represent the input data by the low reconstruction error. The second part is the penalty of sparse coefficient, which is used to constrain the sparsity of coefficients.

2.3 Dictionary Learning

In order to learn the bases (dictionaries) $D=[d_1,...,d_M]\in R^{K\times M}$, the initial dictionary is given which consists of random M samples from the training sets. The process of the iteratively learned dictionary D includes two steps: sparse coding phase and dictionary update phase.

2.3.1 Sparse Coding Phase

Fixing a dictionary D, the best coefficient matrix $\alpha=[\alpha_1,...,\alpha_N]$ can be found. The representation coefficients are computed by orthogonal matching pursuit regression [24]

$$\alpha_n=\arg\min_{\alpha}\left\|x_n-D\alpha\right\|_2^2+\lambda\left\|\alpha_n\right\|_1 \tag{12}$$

2.3.2 Dictionary Update Phase

After the sparse coding is done, a second stage is performed to search for a better dictionary. This process updates one column at a time and the sparse coding dictionary can be updated by optimizing the equation

$$D_n=\arg\min_{D}\frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}\|x_n-D_{n-1}\alpha_n\|_2^2+\lambda\|\alpha_n\|_1 \tag{13}$$

where D_n is the new dictionary, D_{n-1} is the updated dictionary and λ is a regularization parameter.

In this paper, we assume that there are I training classes belonging to I different classes. And $X^i=[x_1,...,x_N]\in R^{K\times N}$ is a speech frame feature sequence of the training samples of the i^{th} object class, where x_n with $n=1,...,N$ is a K dimensional vector stretched by the n^{th} speech frame of the i^{th} object class. $D^i=[d_1,...,d_M]\in R^{K\times M}$ ($K<M$) consisting of M atoms is the dictionary of i^{th} object class in which $d_m=[\mathrm{d}_{1m},...,\mathrm{d}_{Km}]^T$ with $m=1,...,M$ is a K dimensional vector by the m^{th} atom of the i^{th} object class and d_m is a unit column vector. A K-dimension test speech frame $y_s=[\mathrm{y}_{1s},...,\mathrm{y}_{Ks}]^T$ can be well approximated by the linear combination of atoms in the dictionary of the i^{th} object class. For instance, $y_s=D^i\alpha_s^i$ with $\alpha_s^i=[\alpha_{1s}^i,\alpha_{2s}^i,...,\alpha_{Ms}^i]^T$ is the sparse coefficient of the test speech frame y_s in the dictionary D^i of the i^{th} object class. I dictionaries are learned with $i=1,...,I$, where each of them represents a class. The object function of determining D^i is

$$\arg\min_{D^i,\alpha^i}\{\|X^i-D^i\alpha^i\|_2^2+\lambda\|\alpha^i\|_1\}\quad \text{s.t. } d_m^Td_m=1 \tag{14}$$

Where $\alpha^i=[\alpha_1,...,\alpha_N]$ is the representation matrix of X^i over D^i, and the parameter λ is a positive scalar number that balances the l_2 -norm and the l_1 -norm terms. The l_1 penalty yields a sparse solution for α^i of the i^{th} object class, but there is no analytic link between the value of λ and the corresponding effective sparsity $\|\alpha^i\|_0$. In order to prevent D from being arbitrarily large which would lead to arbitrarily small values of α^i, it is common to constrain its column $(d_m)_{m=1}^M$ to have an l_2 norm less than or equal to one. The joint optimization of the dictionary D^i and the coefficients α^i of the sparse decomposition belongs to the NP-hard problem [21, 25], but this problem can be solved when one need be optimized and the other is fixed alternately for D^i and α^i.

2.4 Calculating the Max Pooling Operation on Each Row in Sparse Coding Coefficients

The cortical neuron system encodes the speech in a spike way. When the sound with a particular frequency is transmitted into the neurons, a spike is created at a specific position. While the other neuronal positions keep in the silence state for no response sound incentives. Based on human masking effect, the weak signal is inaudible in the vicinity of a strong signal. A neuron which performs a Max Pooling operation on the pooled inputs responds to the strongest one of its inputs and inhibits other weaker inputs [19]. And according to these character-

istics of the neurons system, the spiking response of the neurons is well predicted by the Max Pooling operation. The Max Pooling operation can be formalized mathematically as returning the largest one of its inputs. In this paper, based on this Max Pooling operation and assuming the dictionary D to be pre-learned, we compute the following speech feature by a pooling function [26]

$$\widetilde{\mathrm{Z}}^{\mathrm{i}} = \psi\left(\alpha^{i}\right) \tag{15}$$

Where α^{i} is the sparse coefficient of the test speech frame sequence Y in the dictionary D^{i} of the i^{th} object class and the pooling function ψ is defined on each row of α^{i}. Each row of α^{i} corresponds to the responses of all the local descriptions to one specific atom in the dictionary D^{i} of the i^{th} object class. We define the pooling function ψ as a Max Pooling function on the absolute sparse codes

$$\mathrm{z}_{\mathrm{m}}^{\mathrm{i}} = \max\{|\ \alpha_{\mathrm{m1}}^{\mathrm{i}}\ |, |\ \alpha_{\mathrm{m2}}^{\mathrm{i}}\ |, \ldots, |\ \alpha_{\mathrm{mN}}^{\mathrm{i}}\ |\} \tag{16}$$

Where, $\mathrm{z}_{\mathrm{m}}^{\mathrm{i}}$ is the m^{th} element of $\widetilde{\mathrm{Z}}^{\mathrm{i}}$ of the i^{th} object class, $\alpha_{\mathrm{mn}}^{\mathrm{i}}$ is the matrix element at the m^{th} row and the n^{th} column of α^{i} with $n = 1,\ldots,N$ and $m = 1,\ldots,M$. Then, $\mathrm{z}_{\mathrm{m}}^{\mathrm{i}}$ of the i^{th} object class is concatenated to form a feature vector representation

$$\widetilde{\mathrm{Z}}^{\mathrm{i}} = [\mathrm{z}_{1}^{\mathrm{i}}, \ldots, \mathrm{z}_{\mathrm{M}}^{\mathrm{i}}]^{\mathrm{T}} \tag{17}$$

where T denotes the transpose operation.

2.5 Computing the Result Score

In the evaluation process, we refer to the method in [15] by selecting J speakers for a speaker verification task, where the speaker selection is random but the claimed speaker is included. For each learned dictionary of J speakers, the test speaker can be represented as a linear combination of the learned dictionary and J groups of sparse coding coefficients are obtained. Then, based on the characteristics of the neurons system for the acoustic stimulus, the Max Pooling operation of J+1 groups are calculated. $\widetilde{Z}^{j}$ ($j \in J$) is used to express the Max Pooling operation in one of J groups of sparse coding coefficients of the j^{th} object class, and $\widetilde{Z}^{claim}$ is used to express the Max Pooling operation of each row element in sparse coding coefficients of the claim object class, where the claimed speaker is coded as a linear combination of his/her claimed dictionary. Since the KL-distance [27] is a natural distance function to measure the similarity of two probabilities, it can be used to compute the similarity between two speaker models. The KL-distance score method is written as,

$$\begin{aligned} R &= \arg\min_{j}\{KL(\widetilde{Z}^{claim}, \widetilde{Z}^{j})\} \\ &= \arg\min_{j}\{\sum_{m} \widetilde{Z}^{claim}(m) \cdot \log_{2} \frac{\widetilde{Z}^{claim}(m)}{\widetilde{Z}^{j}(m)}\} \end{aligned} \tag{18}$$

with $m = 1,\ldots,M$. If the label of R corresponds to the claimed speaker, the verification result indicates the true speaker, otherwise the impostor speaker.

3 Experimental Evaluations

In this study, the MFCC as features is used to represent the speaker characteristic of the train set and the test set. A more compact and individual set of dictionaries is learned from the training samples, where every dictionary is learned from every training sample in the train set and sparse coefficients are obtained. In NIST SRE 2003 speaker recognition evaluation, we select J speaker dictionary models from the training dictionary models where the speaker selection is random but includes the claim speaker. The test feature matrix y can be represented as a linear combination of J speaker dictionary models respectively and obtain J speaker sparse coefficients. Then we calculate the Max Pooling operation on each row in sparse coefficients to concatenate to form the feature vector $\widetilde{Z}^{j}$ ($j \in J$) and $\widetilde{Z}^{claim}$. Finally, the KL-distance function is used to measure the similar score between $\widetilde{Z}^{j}$ ($j \in J$) and $\widetilde{Z}^{claim}$. If the speaker of the similar minimum score corresponds to the claim speaker, the test speaker authenticates the claim speaker. The flow chart of the overview of the proposed method is shown in Fig. 2.

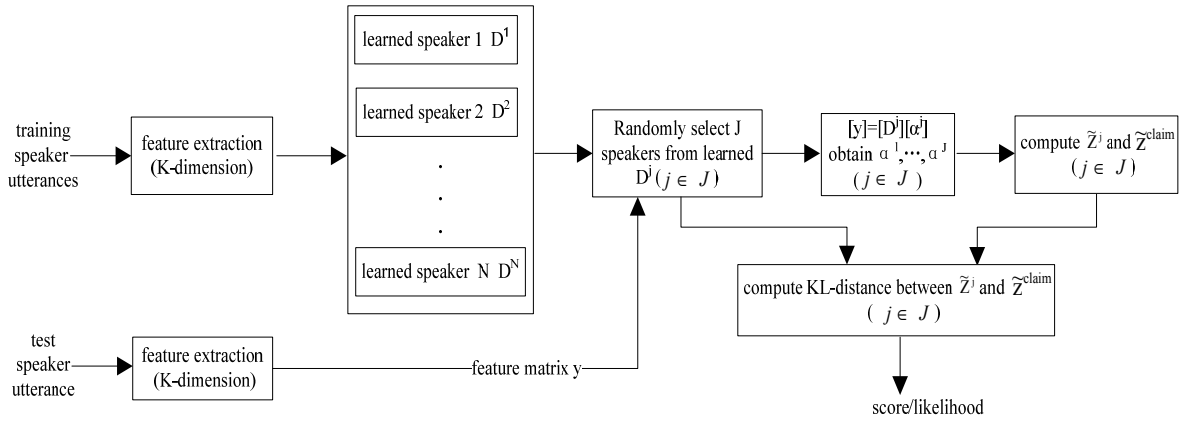

Fig. 2. The flow chart of the overview of the proposed method

In order to assess the discrimination capability of the system, experiments are carried out on the subset of NIST SRE 2003 [28] database and the Chinese-863 database. In NIST-SRE-03, the train set is consisted of 207 target speaker speeches and each of them is approximately 2 minutes. The test set has 1795 true trials and 17950 false trials (I=11). In the Chinese-863 database, the train set is composed of 120 target speaker speeches and each of them is approximately 2 minutes. The test set has 720 true trials and 7200 false trials (I=11). All the speech files are the wave format at the sample frequency of 8 kHz and quantized with 16 bits. Speech streams are windowed into a sequence of short-term frames (20 ms long) with 10 ms overlapped data. Furthermore, the speech files use 34-dimensional MFCC (16+log_energy, appended with their first deltas) with the cepstral mean subtraction (CMS) [6] and the feature warping [6] in order to remove any factors related to the recording conditions.

The baseline system is a gender-dependent universal background model and Gaussian mixture model (GMM-UBM) in which 1024 Gaussians are utilized. In the new feature system, the overcomplete dictionaries contain 256 atoms and each of them is a 34-dimensional vector. The performance of the system is measured by the Equal Error Rate (EER) in which EER is statistical evaluation of the biometric performance of the system. Where, False Acceptance Rate (FAR) and False Rejection Rate (FFR) are equal. In general, the lower the EER is, the more accurate the biometric system is. The results on the subset of NIST SRE 2003 database are shown in Fig. 3, where the solid and dashed lines are used to describe the EERs of the baseline system and the new feature system, respectively. It can be seen that the new feature system (EER with 0.74%) outperforms the baseline system (EER with 10.97%).

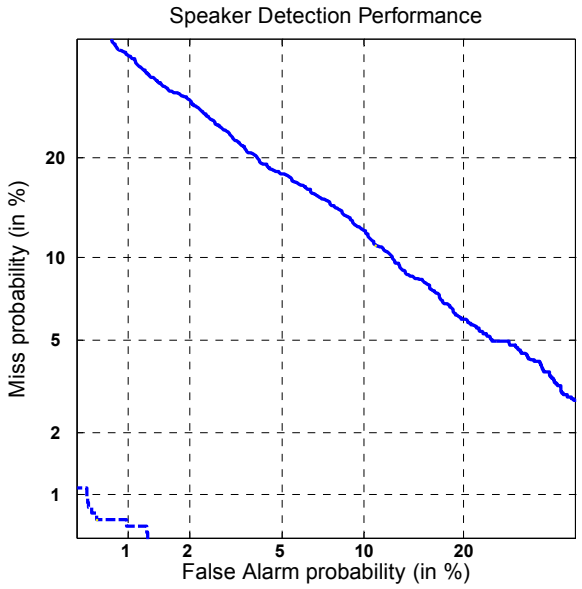

Fig. 3. EERs of the baseline system (the solid line) and the new feature system (the dashed line) in NIST-SRE-03 database.

The results of Chinese-863 database under the noisy environment (0dB) are shown in Fig. 4, where the solid and dashed lines are used to describe the EERs of the baseline system (27.64%) and the new feature system

(14.17%), respectively. The results of Chinese-863 database under the noisy environment (5dB) are shown in Fig. 5. The EER of the baseline system is 16.39% and that of the new system is 11.1%. In the noisy conditions, it can be seen that the new feature system is still better than the baseline system.

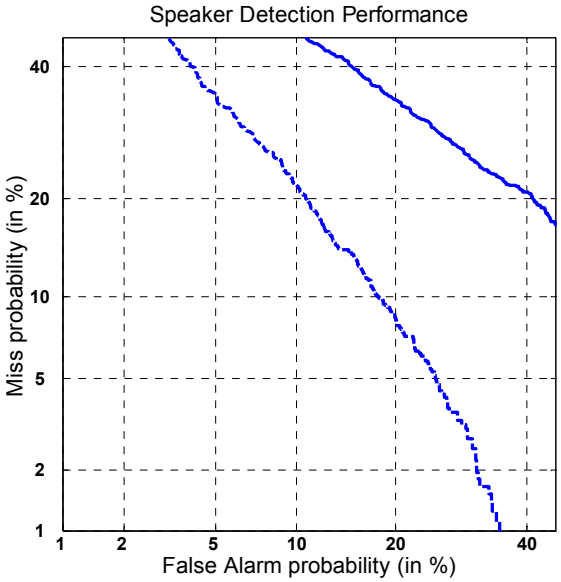

Fig. 4. EERs of the baseline system (the solid line) and the new feature system (the dashed line) in the Chinese-863 database under the noisy environment (0dB).

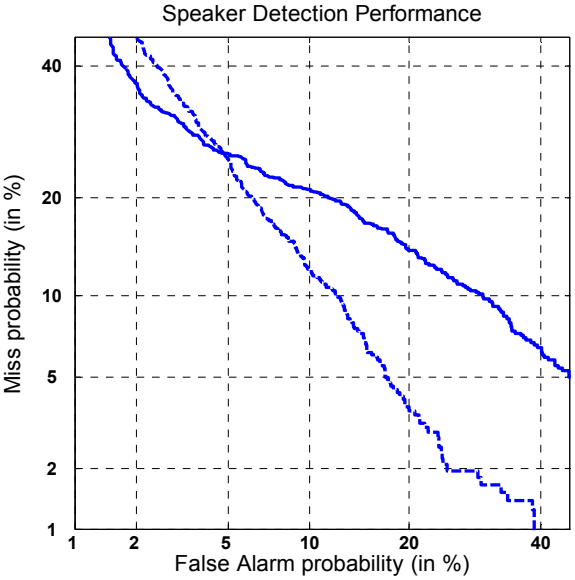

Fig. 5. EERs of the baseline system (the solid line) and the new feature system (the dashed line) in the Chinese-863 database under the noisy environment (5dB).

In order to descript the degree of performance improvement where the results of our proposed system are compared with those of the baseline system, we define relative error rate. Relative error rate is defined as $|EER_{\text{baseline system}} - EER_{\text{proposed system}}| / EER_{\text{baseline system}} * 100\%$. On the subset of NIST SRE 2003 database, relative error rate is 93.25%. On the subset of Chinese-863 2003, relative error rate equals 48.73% under the noisy environment (0dB), and it is 32.28% under the noisy environment (5dB). Such results imply that our approach is strongly robust to noises.

4 Conclusions

In this paper, we propose a biologically inspired feature extraction for speaker verification. According to the acoustic stimulus only activating a small number of cortical neurons in the primary auditory cortex and auditory masking effect, we use the Max Pooling operation of each row in sparse coding coefficients as the feature to express the strongest one of its inputs of every neuronal response (every atom in the dictionary) for the speaker verification. In the ideal and noise environments, the experimental results using this new feature system outperform those of the baseline system. This new feature system will have more important applications in the identity validation fields of the speaker verification such as the entrance guard system, telephone banking and database accessing.

Acknowledgement

This research is supported by the National Natural Science Foundation of China (No. 91120303, No. 91220301) and the Ph.D. programs Foundation of Ministry of Education of China (No. 2011230-2110042).

References

[1] D. Reynolds, "Channel Robust Speaker Verification via Feature Mapping," in *Proceedings of ICASSP 2003*, Vol. 2, pp. II-53-6, 2003.

[2] A. Solomonoff, W. Campbell, I. Boardman, "Advances in Channel Compensation for SVM Speaker Recognition," in *Proceedings of ICASSP 2005*, Vol. 1, pp.629-632, 2005.

[3] S. C. Yin, R. Rose, P. Kenny, "A Joint Factor Analysis Approach to Progressive Model Adaptation in Text-Independent Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, pp. 1999-2010, 2007.

[4] R. Auckenthaler, M. Carey, H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems," *Digital Signal Process*, Vol. 10, No. 1-3, pp. 42-54, 2000.

[5] D. Ramos-Castro, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, J. Ortega-Garcia, "Speaker Verification Using Speaker- and Text-Dependent Fast Score Normalization," *Pattern Recognition*, Vol. 28, No. 1, pp. 90-98, 2007.

[6] T. Kinnunen, H. Li, "An Overview of Text Independent Speaker Recognition: From Features to Supervector," *Speech Communication*, Vol. 52, No. 1, pp. 12-40, 2010.

[7] M. S. Lewicki, "Efficient Coding of Natural Sounds," *Nature Neuroscience*, Vol. 5, pp. 356-363, 2002.

[8] B. A. Olshausen and K. N. O'Connor, "A New Window on Sound," *Nature Neuroscience*, Vol. 5, pp. 292-294, 2002.

[9] K. Huang and S. Aviyente, "Sparse Representation for Signal Classification," *Advances in Neural Information Processing Systems*, Vol. 19, pp. 609, 2007.

[10] S. Mallat, *A Wavelet Tour of Signal Processing*, Second edition, Academic Press, New York, 1999.

[11] E. Candes and F. Guo, "New Multiscale Transforms, Minimum Total Variation Synthesis: Application to Edge Regularization in Image Compression," *Signal Processing*, Vol. 82, No. 11, pp. 1519-1543, 2002.

[12] D. Gabor. "Theory of Communication," *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, Vol. 93, No. 26, pp. 429-457, 1946.

[13] M. Aharon, M. Elad, A. Bruckstein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *IEEE Transactions on Signal Processing*, Vol. 54, pp. 4311-4322, 2006.

[14] J. Mairal, F. Bach, J. Ponce, G. Sapiro, "Online Dictionary Learning for Sparse Coding," in *Proceedings of ICML 2009*,

2009.

[15] J. M. K. Kua, "Speaker Verification Using Sparse Representation Classification," in *Proceedings of ICASSP 2011*, pp. 4548-4551, 2011.

[16] T. N. Sainath, A. Carmi, D. Kanevsky, B. Ramabhadran, "Bayesian Compressive Sensing for Phonetic Classification," in *Proceedings of ICASSP 2010*, pp. 4370-4373, 2010.

[17] S. F. Cotter, "Sparse Representation for Accurate Classification of Corrupted and Occluded Facial Expressions," in *Proceedings of ICASSP 2010*, pp. 838-841, 2010.

[18] B. C. Haris, R. Sinha, "Exploring Sparse Representation Classification for Speaker Verification in Realistic Environment," in *Proceedings of Centenary Conference*, IISc Bangalore, 2011.

[19] H. Fletcher, "A Space-Time Pattern Theory of Hearing," *Journal of the Acoustical Society of America*, Vol. 1, No. 3A, pp. 311-343, 1930.

[20] M. Li, X. Zhang, Y. Yan, S. Narayanan, "Speaker Verification Using Sparse Representations on Total Variability i-Vectors," in *Proceedings of Interspeech 2011*, pp. 2729-2732, 2011.

[21] M. Aharon, M. Elad, A. M. Bruckstein, "The KSVD: An Algorithm for Designing of Overcomplete Dictionaries for Sparse Representations," *IEEE Transactions Signal Processing*, Vol. 54, pp. 4311-4322, 2006.

[22] S. A. Gelfand, *Hearing: An Introduction to Psychological and Physiological Acoustics*, 4th Ed., New York, Marcel Dekker, 2004.

[23] B. A. Olshausen, D. J. Field, "Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images," *Nature*, Vol. 381, pp. 607-609, 1996.

[24] Y. C. Pati, R. Rezaiifar, P. S. Krishnaprasad, "Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition," in *Proceedings of 27th Annual Asilomar Conf. on Signal, Systems, and Computers*, 1993

[25] H. Lee, A. Battle, R. Raina, A. Y. Ng, "Efficient Sparse Coding Algorithms," *Advances in Neural Information Processing Systems*, Vol. 19, pp. 801-808, 2007.

[26] J. Yang, K. Yu, Y. Gong, T. S. Huang, "Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification," in *Proceedings of CVPR 2009*, 2009.

[27] P. J. Moreno, P. P. Ho, N. Vasconcelos, "A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications," in *Proceedings of the Advances in Neural Information Processing Systems 2004*, Cambridge, MIT Press, MA, Vol. 16, 1385-1392, 2004.

[28] M. Przybocki and A. Martin, The NIST Year 2003 Speaker Recognition Evaluation Plan, http://www.nist.gov/tests/spk/2003/index.htm, 2003.