Robust Auditory-Based Speech Feature Extraction Using Independent Subspace Method

Qiang Wu[1], Liqing Zhang[1], and Bin Xia[2]

[1] Department of Computer Science and Engineering,
Shanghai Jiao Tong University, Shanghai 200240, China
[2] Department of Electronic Engineering,
Shanghai Maritime University, Shanghai 200135, China
{johnnywu,lqzhang}@sjtu.edu.cn, binxia@cie.shmtu.edu.cn

Abstract. In recent years many approaches have been developed to address the problem of robust speaker recognition in adverse acoustical environments. In this paper we propose a robust auditory-based feature extraction method for speaker recognition according to the characteristics of the auditory periphery and cochlear nucleus. First, speech signals are represented based on frequency selectivity at basilar membrane and inner hair cells. Then, features are mapped into different linear subspaces using independent subspace analysis (ISA) method, which can represent some high order, invariant statistical features by maximizing the independence between norms of projections. Experiment results demonstrate that our method can considerably increase the speaker recognition accuracy specifically in noisy environments.

1 Introduction

In speaker recognition system, feature extraction is one of important tasks, which aims at finding succinct, robust, and discriminative features from acoustic data. Acoustic features such as linear predictive cepstral coefficients (LPCC)[1], mel-frequency cepstral coefficients (MFCC)[1], perceptual linear predictive coefficients (PLP)[1] are commonly used, and the most popular data modeling techniques in current speaker recognition are based on the gaussian mixture model (GMM)[2]. Recently the computational auditory nerve models attract much attention from both neuroscience and speech signal processing communities. Lewicki et al.[3] demonstrated that efficient coding of natural sounds could explain auditory nerve filtering properties and their organization as a population. Smith et al.[4] proposed an algorithm for learning efficient auditory codes using a theoretical model for coding sound in terms of spikes.

However, the conventional feature extraction methods for speaker recognition are often affected by the environmental noise or channel distortions. In this paper, we investigate statistical approaches of constructing a basis function for encoding patterns including spectral and temporal information. This method attempts to extract the robust speech features by mapping the frequency selectivity characteristics at cochlea into independent subspace. The extraction features may model the differences of speakers and reduce the disturbances of noise. Furthermore, we employ the support vector machine as a classifier to test the effectiveness and recognition performance.

2 Method

As we know, the human auditory system can accomplish the speaker recognition easily and be insensitive to the background noise. In our approach, the first step is to obtain the frequency selectivity information by imitating the process performed in the auditory periphery and pathway. And then we represent the robust speech feature as the projections of the extracted auditory information mapped into a feature subspace via independent subspace analysis. A diagram of the feature extraction method and speaker recognition is shown in Figure.1. For the auditory-based processing of speech signals, we imple-

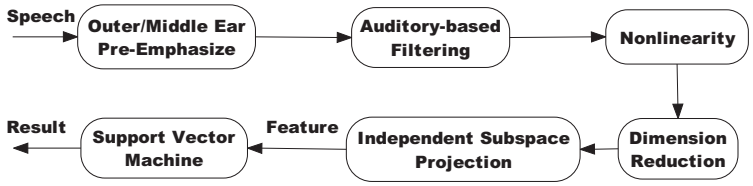


Fig. 1. Extraction of auditory feature by independent subspace method and recognition

ment three modules of auditory system as what is described in Figure.1 to obtain the representation in the auditory periphery and pathway .

In order to raise the energy for frequency components located in the high frequency domain, we implement traditional pre-emphasis to model the combined outer and middle ear as a band-pass function i.e. $H(z) = 1 - 0.97z^{-1}$.

The frequency selectivity of peripheral auditory system is simulated by a bank of cochlear filters which represent frequency selectivity at various locations along the basilar membrane in a cochlea. The cochlear filterbanks have an impulse response as $g(t) = at^{n-1}e^{2\pi bERB(f_c)t}cos(2\pi f_c t + \phi)$, where n is the order of the filter, f_c is the center frequency, ϕ is the phase, $a, b \in R$ are constants where b determines the rate of decay of the impulse response, which is related to bandwidth, $ERB(f_c)$ is the equivalent rectanglar bandwidth (ERB) of the auditory filter with a quadratic formula i.e. $ERB = 24.7(4.37f_c/1000 + 1)$.

In order to model nonlinearity of the inner hair-cells, the power of each band in every frame with a logarithmic nonlinearity was calculated by following equation i.e. $x(k) = log(1+\gamma \sum_{t \in frame\, k}\{x_g(t)\}^2)$, where $x(k)$ is the output power, γ is a scaling constant, $x_g(t)$ is the output of cochlear filterbanks. This result can be considered as average firing rates in the inner hair-cells, which simulate the higher auditory pathway.

In this paper we apply independent subspace analysis to learn an optimal basis function which can give a robust representation of speech signals. The motivation of independent subspace analysis[5] (ISA) is to achieve such an extension which generalizes the assumption of component independence to independence between groups of components. Compared to the ordinary ICA model, the components s_i in ISA model are not assumed to be all mutually independent. Instead, s_i can be divided into n-tuples and the s_i inside a given n-tuple may be dependent on each other, but dependencies among dif-

ferent n-tuples are not allowed. A stochastic gradient ascent algorithm can be described as:

$$\Delta \mathbf{w}_i \propto \mathbf{x}(\mathbf{w}_i^T \mathbf{x}) g(\sum_{i \in S_j} (\mathbf{w}_i^T \mathbf{x})^2). \quad (1)$$

where $\mathbf{w}_i$ is the vector of demixing matrix, $\mathbf{x}$ is the observed signal, $g = p'/p$ is a nonlinear function that incorporates the information on the sparseness of the norms of the projections. In order to speed up the convergence, we prewhite the signals and constrain the vectors $\mathbf{w}_i$ to be orthogonal and unit norm. More information about ISA can be found in [5].

3 Experiments and results

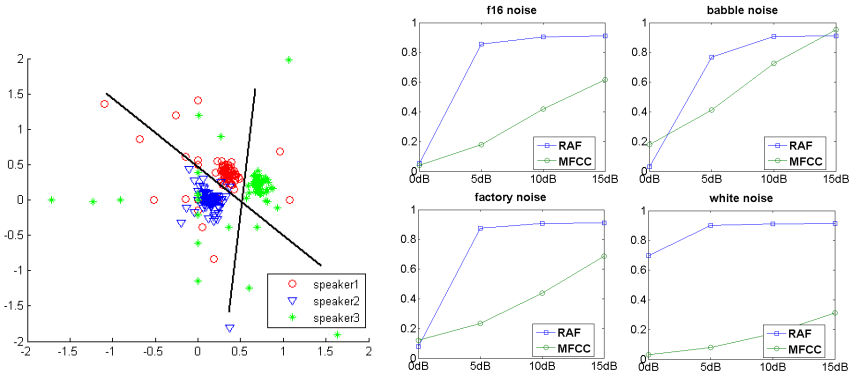
In order to evaluate the efficiency of our method, a text-independent speaker identification experiment was conducted. We used Grid speech corpus to test the performance of our feature extraction method in section 2. The Grid speech Corpus contains 17000 sentences spoken by 34 speakers(18 males and 16 females).

In our experiments the sampling rate of speech signals was 8kHz. For the given speech signals, we employed every window of length 8000 samples(1s) and time duration 20 samples(2.5ms) and 36 gammatone filters were selected. In order to reduce the computational complexity, principle component analysis was performed for the dimension reduction. As described in Section.2, we calculated the basis function using independent subspace analysis after the calculation of the average firing rates in the inner hair-cells. 170 sentences(5 sentences each person) were selected randomly as the training data for learning basis function and 40 independent feature subspaces were obtained which subspace dimension was chosen to be 4.

In order to test the efficiency and robustness of our feature extraction method, we employed support vector machine as the classifier. 1700 sentences (50 sentences each person) were used as training data and 2040 sentences (60 sentences each person) mixed with different kinds of noise were used as test data. The test data was mixed with babble, factory, f16 and white noises in SNR intensities of 15dB, 10dB, 5dB and 0dB.

For comparison, we implemented a baseline GMM system that used conventional MFCC. In the system, each frame was modeled by 13-component vector, derived from a 40-channel Mel-scale filter bank, and the popular data modeling method GMM was used to build the recognizer with 32 gaussian mixtures. From Figure.2(a) we can observe the classification boundaries among different speakers clearly which is beneficial to the identification. Figure.2(b) presents the identification accuracy obtained by the robust auditory-based feature (RAF) and baseline system in all tested conditions.

The result demonstrates that the performance degradation of RAF is little with noise intensity increase. It performs significantly better than MFCC in the high noise conditions. Such as 0dB condition of white noise, RAF maintains results close to the low noise condition, while MFCC has dropped sharply. The result can suggest that this auditory-based method is robust against the noise and improves the recognition performance.



(a) (b)

Fig. 2. Feature space for classification and identification accuracy. (a) depicts projections onto two-dimensional feature space using PCA. The data was mixed with white noise in SNR 5dB. (b) shows the identification accuracy in different noise conditions.

4 Conclusions

An auditory-based feature extraction method applied to a text-independent speaker recognition task was presented in this paper, and according to the experiments results, the robustness and effectiveness were confirmed. This method is designed to extract the robust speech features by mapping the frequency selectivity characteristics at cochlea into independent subspace by learning a basis function. The goal of finding a optimal basis function using independent subspace analysis was to increase the robustness of feature by removing the noisy components and improve the recognition scores.

Acknowledgment

The work was supported by the National High-Tech Research Program of China (Grant No.2006AA01Z125) and the National Basic Research Program of China (Grant No. 2005CB724301).

References

1. L.R.Rabiner, B.Juang: Fundamentals on Speech Recognition. Prentice Hall, New Jersey (1996)
2. D.A.Reynolds, R.C.Rose: Robust text-independent speaker identification using gaussianmixture speaker models. Speech and Audio Processing, IEEE Transactions on **3** (1995) 72–83
3. M.S.Lewicki: Efficient coding of natural sounds. Nature neuroscience **5** (2002) 356–363
4. E.C.Smith, M.S.Lewicki: Efficient auditory coding. Nature **439** (2006) 978–982
5. A.Hyv arinen, P.Hoyer: Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. Neural Computation **12** (2000) 1705–1720