Wu Q, Zhang LQ, Shi GC. Robust feature extraction for speaker recognition based on constrained nonnegative tensor factorization. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 25(4): 745–754 July 2010/DOI 10.1007/s11390-010-1061-z

Robust Feature Extraction for Speaker Recognition Based on Constrained Nonnegative Tensor Factorization

Qiang Wu (吴　强), Li-Qing Zhang* (张丽清), and Guang-Chuan Shi (石光川)

Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

E-mail: {johnnywu, lqzhang, sgc1984}@sjtu.edu.cn

Received May 6, 2009; revised January 24, 2010.

Abstract How to extract robust feature is an important research topic in machine learning community. In this paper, we investigate robust feature extraction for speech signal based on tensor structure and develop a new method called constrained Nonnegative Tensor Factorization (cNTF). A novel feature extraction framework based on the cortical representation in primary auditory cortex (A1) is proposed for robust speaker recognition. Motivated by the neural firing rates model in A1, the speech signal first is represented as a general higher order tensor. cNTF is used to learn the basis functions from multiple interrelated feature subspaces and find a robust sparse representation for speech signal. Computer simulations are given to evaluate the performance of our method and comparisons with existing speaker recognition methods are also provided. The experimental results demonstrate that the proposed method achieves higher recognition accuracy in noisy environment.

Keywords pattern recognition, speaker recognition, nonnegative tensor factorization, feature extraction, auditory perception

1 Introduction

Finding succinct, robust and discriminative features from acoustic data is one of important tasks for a speaker recognition system. Acoustic features such as Linear Prediction Cepstral Coefficients (LPCC), Mel-Frequency Cepstral Coefficients (MFCC)[1], Perceptual Linear Predictive (PLP)[2] are common solutions. And conventional speaker modeling methods such as Gaussian mixture models (GMM)[3] achieve very high performance for speaker identification and verification tasks on high-quality data when training and testing conditions are well controlled.

However, in the real application such systems usually do not perform well due to a large variety of speech signals corrupted by adverse conditions such as environmental noise and channel distortions. Numerous efforts have been made on the robust speech feature extraction to adapt noisy environments. Feature compensation techniques[4-7] such as cepstral mean normalization (CMN), RASTA, have been developed for robust speech recognition. Spectral subtraction[8-9] and subspace-based filtering[10-11] techniques have been widely used because of their simplicity. Dimension reduction methods such as PCA, LDA and NMF[12-13] are also proved to be efficient for speech feature extraction in noisy environment. These methods bring certain process, but the overall performances are still far from satisfactory. Some methods need noise information beforehand. Non-stationary noise and low signal-to-noise ratio (SNR) are still open problems.

Human auditory system can solve all these problems quite well. Recently, many researchers explore how to incorporate models of auditory system for robust speech feature extraction. In the procedure of the peripheral auditory system, the auditory spectrum is encoded by population of cortical cells in primary auditory cortex and separated into different cues and features associated with different sound percepts. In the central auditory system, the speech spectrum can be analyzed as spectral and temporal modulations by a bank modulation-selective filters in primary auditory cortex[14]. An analytically tractable framework based on wavelet representation and multiresolution processing[15] was described to explain mechanical and neural processing in the early stages of the auditory system. Mesgarani[16] investigated audio classification problem based on multiscale spectro-temporal modulation features inspired

Regular Paper

The work was supported by the National Natural Science Foundation of China under Grant No. 60775007, the National Basic Research 973 Program of China under Grant No. 2005CB724301, and the Science and Technology Commission of Shanghai Municipality under Grant No. 08511501701.

*Corresponding Author

by auditory cortex model and applied HOSVD to perform multilinear dimensionality reduction. Jeon[17] proposed a computational auditory central system model based on the physiological model in [14]. This model is a data-redundant, multilinear representation of neural firing rates in A1 and has been validated in a conventional phoneme classification task.

All these feature extraction or denoising methods investigate the representation in time or spectro-temporal domain, while for speech feature extraction especially in the noisy condition we may need more information or factors to improve the robustness of features. Multi-factors analysis has been brought into consideration for speech feature extraction or general audio signal processing tasks. Currently there are two types of common tensor decomposition methods such as CANDECOMP/PARAFAC model[18-20] and Tucker Model[21-23]. Nonnegative Tensor Factorization (NTF)[24-26] imposes the nonnegative constraint on the CANDECOMP/PARAFAC model. In computer vision, tensor factorization approach was used as image representation. Vasilescu[27] introduced a multifactor model called Multilinear ICA to learn the statistically independent components of multiple factors resulting in a better performance for face recognition. Tao[28-29] developed general tensor discriminant analysis for the gait recognition which reduces the under sample problem.

In this paper, we propose a new framework for robust speaker modeling based on cortical model and tensor factorization. Firstly, we investigate the cortical representation of speech signal in the primary auditory cortex, which is a intrinsic array structure with multiple factors. Then, a new tensor factorization method with orthogonal and nonnegative constraints called cNTF is developed to learn the basis functions of multi-related feature subspaces from the cortical representation. Sparse constraint on basis functions enhances energy concentration of speech signal, keeping the useful feature during the noise reduction. The features extracted by cNTF can be further processed into a representation called Cortical Tensor Cepstral Coefficients (CTCC) via discrete cosine transform. The CTCC is used as feature representation for speaker recognition in this paper. Finally, GMM is employed to perform speaker modeling and recognition.

The reminder of this paper is organized as follows. In Section 2 a new supervised learning algorithm cNTF is derived for feature extraction. Section 3 describes the auditory model and sparse tensor feature extraction framework. Section 4 presents the experimental results for speaker identification task on Aurora2 dataset recorded in noisy environments. Finally, Section 5 gives a conclusion of this paper.

2 Method

As an extension of matrix factorization, PARAFAC model is an efficient tool for high order data analysis. In this section, a new tensor factorization algorithm called cNTF is proposed to extend PARAFAC model by orthogonal and nonnegative constraints. First, a brief overview of multilinear algebra and PARAFAC model is given. Then cNTF algorithm is presented. Some basic notations of multilinear algebra are described in Table 1.

Table 1. Notations in Multilinear Algebra

Notation	Description
$\mathcal{X}$	N-way tensor
$\boldsymbol{X}$	Matrix
$\boldsymbol{X}_{(d)}$	Mode-d matricization of tensor $\mathcal{X}$
$\odot$	Khatri-Rao product
$\circ$	Outer product
$\otimes$	Kronecker product

2.1 PARAFAC Model

Multilinear algebra is the algebra of higher order tensors. A tensor is a higher order generalization of a matrix. Let $\mathcal{X} \in \mathbb{R}^{N_1 \times N_2 \times \cdots \times N_M}$ denote a tensor of order M. An element of $\mathcal{X}$ is denoted by $x_{n_1,n_2,\ldots,n_M}$, where $1 \leqslant n_d \leqslant N_d$ and $1 \leqslant d \leqslant M$. The mode-$d$ matricization or matrix unfolding of an Mth-order tensor $\mathcal{X} \in \mathbb{R}^{N_1 \times N_2 \times \cdots \times N_M}$ rearranges the elements of $\mathcal{X}$ to form the matrix $\boldsymbol{X}_{(d)} \in \mathbb{R}^{N_d \times N_{\bar{d}}}$, which is the ensemble of vectors in $\mathbb{R}^{N_d}$ obtained by keeping index n_d fixed and varying the other indices. Here $N_{\bar{d}} = \prod_{j \neq d}^{M} N_j$. Matricizing a tensor is similar to vectoring a matrix.

The PARAFAC model was suggested independently by Carroll and Chang[18] under the name CANDECOMP (canonical decomposition) and by Harshman[19] under the name PARAFAC (parallel factor analysis) which has gained increasing attention in data mining fields. This model has structural resemblance with many physical models of common real-world data and its uniqueness property[30] implies that data following the PARAFAC model can be uniquely decomposed into individual components.

An M-way tensor $\mathcal{X} \in \mathbb{R}^{N_1 \times N_2 \times \cdots \times N_M}$ can be decomposed in a sum of M rank-1 terms, i.e., represented by the outer product of M vectors[31]:

$$\mathcal{X} = \sum_{r=1}^{R} \boldsymbol{A}_{:,r}^{(1)} \circ \boldsymbol{A}_{:,r}^{(2)} \circ \cdots \circ \boldsymbol{A}_{:,r}^{(M)}, \tag{1}$$

where $\boldsymbol{A}_{:,r}^{(d)}$ represents the r-th column vector of the mode matrix $\boldsymbol{A}^{(d)} \in \mathbb{R}^{N_d \times R}$, the rank of tensor $\mathcal{X}$, denoted by $R = rank(\mathcal{X})$, is the minimal number of rank-1 tensors.

Given a tensor $\mathcal{X} \in \mathbb{R}^{N_1 \times N_2 \times \cdots \times N_M}$, PARAFAC model aims to find a rank-R approximation of the tensor $\mathcal{X}$ in the form of (1),

$$\mathcal{X} \approx \sum_{r=1}^{R} \boldsymbol{A}^{(1)}_{:,r} \circ \boldsymbol{A}^{(2)}_{:,r} \circ \cdots \circ \boldsymbol{A}^{(M)}_{:,r}, \tag{2}$$

or in the element-wise form

$$x_{n_1,n_2,\ldots,n_M} \approx \sum_{r=1}^{R} a^{(1)}_{n_1 r} a^{(2)}_{n_2 r} \cdots a^{(M)}_{n_M r}. \tag{3}$$

The PARAFAC model can also be written in matrix notation by use of the Khatri-Rao product, which gives the equivalent expressions:

$$\boldsymbol{X}_{(d)} \approx \boldsymbol{A}^{(d)}[\boldsymbol{A}^{(d-1)} \odot \cdots \odot \boldsymbol{A}^{(1)} \odot \boldsymbol{A}^{(M)} \odot \cdots \odot \boldsymbol{A}^{(d+1)}]^{\mathrm{T}}. \tag{4}$$

For further detailed discussions about tensor factorization, refer to [31].

2.2 Constrained Nonnegative Tensor Factorization

Given a nonnegative M-way tensor $\mathcal{X} \in \mathbb{R}^{N_1 \times N_2 \times \cdots \times N_M}$, nonnegative tensor factorization (NTF) seeks a factorization of $\mathcal{X}$ in the form:

$$\mathcal{X} \approx \hat{\mathcal{X}} = \sum_{r=1}^{R} \boldsymbol{A}^{(1)}_{:,r} \circ \boldsymbol{A}^{(2)}_{:,r} \circ \cdots \circ \boldsymbol{A}^{(M)}_{:,r}, \tag{5}$$

where the mode matrices $\boldsymbol{A}^{(d)} \in \mathbb{R}^{N_d \times R}$ for $d = 1, \ldots, M$ are restricted to have only nonnegative elements in the factorization. In order to find an approximate tensor factorization $\hat{\mathcal{X}}$, we can construct Least Square cost function $\boldsymbol{J}_{LS}$ and KL-divergence cost function $\boldsymbol{J}_{KL}$ as follows

$$\begin{aligned}\boldsymbol{J}_{LS}(\mathcal{X}, \hat{\mathcal{X}}) &= \frac{1}{2}\|\mathcal{X} - \hat{\mathcal{X}}\|_F^2 \\ &= \frac{1}{2} \sum_{n_1,n_2,\cdots,n_M} \left(x_{n_1,n_2,\ldots,n_M} - \sum_{r=1}^{R} a^{(1)}_{n_1 r} a^{(2)}_{n_2 r} \cdots a^{(M)}_{n_M r} \right)^2,\end{aligned} \tag{6}$$

$$\begin{aligned}\boldsymbol{J}_{KL}(\mathcal{X}, \hat{\mathcal{X}}) &= D(\mathcal{X}\|\hat{\mathcal{X}}) \\ &= \sum_{n_1,n_2,\ldots,n_M} \left(x_{n_1,n_2,\ldots,n_M} \log \frac{x_{n_1,n_2,\ldots,n_M}}{\hat{x}_{n_1,n_2,\ldots,n_M}} - \right. \\ &\quad \left. x_{n_1,n_2,\ldots,n_M} + \hat{x}_{n_1,n_2,\ldots,n_M} \right)\end{aligned} \tag{7}$$

where $\|\cdot\|_F^2$ is the Frobenius norm.

Based on the approximate factorization model (4), we can redefine the cost function with mode matrices $\boldsymbol{A}^{(d)}$

$$\begin{aligned}\boldsymbol{J}_{LS1}(\boldsymbol{A}^{(d)}|_{d=1}^{M}) &= \frac{1}{2}\sum_{d=1}^{M} \|\boldsymbol{X}_{(d)} - \boldsymbol{A}^{(d)}\boldsymbol{Z}^{(d)}\|_F^2 \\ &= \frac{1}{2}\sum_{d=1}^{M}\sum_{p=1}^{N_d}\sum_{q=1}^{N_{\bar{d}}}([\boldsymbol{X}_{(d)}]_{pq} - [\boldsymbol{A}^{(d)}\boldsymbol{Z}^{(d)}]_{pq})^2,\end{aligned} \tag{8}$$

$$\begin{aligned}\boldsymbol{J}_{KL1}(\boldsymbol{A}^{(d)}|_{d=1}^{M}) &= \sum_{d=1}^{M} D(\boldsymbol{X}_{(d)}\|\boldsymbol{A}^{(d)}\boldsymbol{Z}^{(d)}) \\ &= \sum_{d=1}^{M}\sum_{p=1}^{N_d}\sum_{q=1}^{N_{\bar{d}}} \Big([\boldsymbol{X}_{(d)}]_{pq} \log \frac{[\boldsymbol{X}_{(d)}]_{pq}}{[\boldsymbol{A}^{(d)}\boldsymbol{Z}^{(d)}]_{pq}} - \\ &\quad [\boldsymbol{X}_{(d)}]_{pq} + [\boldsymbol{A}^{(d)}\boldsymbol{Z}^{(d)}]_{pq} \Big),\end{aligned} \tag{9}$$

where

$$\boldsymbol{Z}^{(d)} = [\boldsymbol{A}^{(d-1)} \odot \cdots \odot \boldsymbol{A}^{(1)} \odot \boldsymbol{A}^{(M)} \odot \cdots \odot \boldsymbol{A}^{(d+1)}]^{\mathrm{T}} \tag{10}$$

and $N_{\bar{d}} = \prod_{j \neq d}^{M} N_j$. These cost functions are quite similar to NMF[32], which perform matrix factorization in each mode and minimize the error for all modes. In this model, we can impose additional constraint which makes the basis functions be as orthogonal as possible, i.e., ensures redundancy minimization between different basis functions. Then, the cost functions become:

$$\begin{aligned}&\boldsymbol{J}_{LS2}(\boldsymbol{A}^{(d)}|_{d=1}^{M}) \\ &= \sum_{d=1}^{M} \Big(\frac{1}{2}\sum_{p=1}^{N_d}\sum_{q=1}^{N_{\bar{d}}}([\boldsymbol{X}_{(d)}]_{pq} - [\boldsymbol{A}^{(d)}\boldsymbol{Z}^{(d)}]_{pq})^2 + \\ &\quad \alpha \sum_{p \neq q} [\boldsymbol{A}^{(d)T}\boldsymbol{A}^{(d)}]_{pq} \Big),\end{aligned} \tag{11}$$

$$\begin{aligned}&\boldsymbol{J}_{KL2}(\boldsymbol{A}^{(d)}|_{d=1}^{M}) \\ &= \sum_{d=1}^{M} \Big(\sum_{p=1}^{N_d}\sum_{q=1}^{N_{\bar{d}}} \Big([\boldsymbol{X}_{(d)}]_{pq} \log \frac{[\boldsymbol{X}_{(d)}]_{pq}}{[\boldsymbol{A}^{(d)}\boldsymbol{Z}^{(d)}]_{pq}} - \\ &\quad [\boldsymbol{X}_{(d)}]_{pq} + [\boldsymbol{A}^{(d)}\boldsymbol{Z}^{(d)}]_{pq}\Big) + \alpha \sum_{p \neq q} [\boldsymbol{A}^{(d)T}\boldsymbol{A}^{(d)}]_{pq} \Big),\end{aligned} \tag{12}$$

where α is a balancing parameter between reconstruction and orthogonality.

A number of approaches have been proposed to control the sparseness by additional constraints or penalization terms. These constraints or penalizations can be applied to the basis vectors or both basis and encoding vectors. The nonsmooth NMF (nsNMF) model[33] proposed a factorization model $\boldsymbol{V} = \boldsymbol{WSH}$, providing a smoothing matrix $\boldsymbol{S} \in \mathbb{R}^{k \times k}$ given by

$$\boldsymbol{S} = (1 - \theta)\boldsymbol{I} + \frac{\theta}{k}\mathbf{1}\mathbf{1}^{\mathrm{T}}, \tag{13}$$

where $\|\cdot\|^T$ is the transpose operater, $\boldsymbol{I}$ is the identify matrix, $\boldsymbol{1}$ is a vector of ones, and the parameter θ satisfies $0 \leqslant \theta \leqslant 1$. For $\theta = 0$, the model (13) is equivalent to the original NMF. As $\theta \to 1$, stronger smoothing is imposed on $\boldsymbol{S}$, leading to a strong sparseness on both $\boldsymbol{W}$ and $\boldsymbol{H}$. By this nonsmooth approach, we can control the sparseness of basis vectors and encoding vectors and maintain the faithfulness of the model to the data. The corresponding cost functions can be given by

$$\begin{aligned}&\boldsymbol{J}_{LS3}(\boldsymbol{A}^{(d)}|_{d=1}^{M})\\&=\sum_{d=1}^{M}\Big(\frac{1}{2}\sum_{p=1}^{N_d}\sum_{q=1}^{N_{\bar{d}}}([\boldsymbol{X}_{(d)}]_{pq}-[\boldsymbol{A}^{(d)}\boldsymbol{S}\boldsymbol{Z}^{(d)}]_{pq})^2+\\&\quad\alpha\sum_{p\neq q}[\boldsymbol{A}^{(d)\mathrm{T}}\boldsymbol{A}^{(d)}]_{pq}\Big),\end{aligned}\tag{14}$$

$$\begin{aligned}&\boldsymbol{J}_{KL3}(\boldsymbol{A}^{(d)}|_{d=1}^{M})\\&=\sum_{d=1}^{M}\Big(\sum_{p=1}^{N_d}\sum_{q=1}^{N_{\bar{d}}}\Big([\boldsymbol{X}_{(d)}]_{pq}\log\frac{[\boldsymbol{X}_{(d)}]_{pq}}{[\boldsymbol{A}^{(d)}\boldsymbol{S}\boldsymbol{Z}^{(d)}]_{pq}}-\\&\quad[\boldsymbol{X}_{(d)}]_{pq}+[\boldsymbol{A}^{(d)}\boldsymbol{S}\boldsymbol{Z}^{(d)}]_{pq})+\alpha\sum_{p\neq q}[\boldsymbol{A}^{(d)\mathrm{T}}\boldsymbol{A}^{(d)}]_{pq}\Big).\end{aligned}\tag{15}$$

The gradient for $\boldsymbol{A}_{ij}^{(d)}$ can be obtained:

$\boldsymbol{LS}$:

$$\begin{aligned}&\frac{\partial\boldsymbol{J}_{LS3}(\boldsymbol{A}^{(d)})}{\partial\boldsymbol{A}_{ij}^{(d)}}=-\big([\boldsymbol{X}_{(d)}\boldsymbol{Z}^{(d)\mathrm{T}}\boldsymbol{S}^{\mathrm{T}}]_{ij}-\\&\quad[\boldsymbol{A}_{ij}^{(d)}\boldsymbol{S}\boldsymbol{Z}^{(d)}\boldsymbol{Z}^{(d)T}\boldsymbol{S}^{T}]_{ij}\big)+\alpha\sum_{p\neq j}[\boldsymbol{A}^{(d)\mathrm{T}}]_{pi},\end{aligned}\tag{16}$$

$\boldsymbol{KL}$:

$$\begin{aligned}&\frac{\partial\boldsymbol{J}_{KL3}(\boldsymbol{A}^{(d)})}{\partial\boldsymbol{A}_{ij}^{(d)}}=-\Big(\sum_{k}[\boldsymbol{S}\boldsymbol{Z}^{(d)}]_{jk}\frac{[\boldsymbol{X}_{(d)}]_{ik}}{[\boldsymbol{A}^{(d)}\boldsymbol{S}\boldsymbol{Z}^{(d)}]_{ik}}-\\&\quad\sum_{k}[\boldsymbol{S}\boldsymbol{Z}^{(d)}]_{jk}\Big)+\alpha\sum_{p\neq j}[\boldsymbol{A}^{(d)\mathrm{T}}]_{pi}.\end{aligned}\tag{17}$$

Here we derive multiplicative learning algorithms for mode matrices $\boldsymbol{A}^{(d)}$ using the Exponential Gradient, which are similar to those in NMF. The monotonic convergence analysis in [32] can be applied to our case as well. Updating algorithms in an element-wise manner for minimizing the cost function (14) and (15) are directly derived as done in [24-25]:

$\boldsymbol{LS}$:

$$\boldsymbol{A}_{ij}^{(d)}\leftarrow\frac{\boldsymbol{A}_{ij}^{(d)}[\boldsymbol{X}_{(d)}\boldsymbol{Z}^{(d)\mathrm{T}}\boldsymbol{S}^{\mathrm{T}}]_{ij}}{[\boldsymbol{A}_{ij}^{(d)}\boldsymbol{S}\boldsymbol{Z}^{(d)}\boldsymbol{Z}^{(d)\mathrm{T}}\boldsymbol{S}^{\mathrm{T}}]_{ij}+\alpha\sum_{p\neq j}[\boldsymbol{A}^{(d)\mathrm{T}}]_{pi}},\tag{18}$$

$\boldsymbol{KL}$:

$$\boldsymbol{A}_{ij}^{(d)}\leftarrow\frac{\boldsymbol{A}_{ij}^{(d)}\sum_{k}[\boldsymbol{S}\boldsymbol{Z}^{(d)}]_{jk}\frac{[\boldsymbol{X}_{(d)}]_{ik}}{[\boldsymbol{A}^{(d)}\boldsymbol{S}\boldsymbol{Z}^{(d)}]_{ik}}}{\sum_{k}[\boldsymbol{S}\boldsymbol{Z}^{(d)}]_{jk}+\alpha\sum_{p\neq j}[\boldsymbol{A}^{(d)\mathrm{T}}]_{pi}}.\tag{19}$$

Table 2 lists the alternating projection optimization procedure for constraints Nonnegative Tensor Factorization. The key steps in the alternating projection procedure are step 3, which aims to find the basis functions $[\boldsymbol{A}^{(d)}]^{(t)}$ in the t-th iteration by using $[\boldsymbol{A}^{(d)}]^{(t-1)}|_{d=1}^{M}$ found in the $(t-1)$-th iteration.

Table 2. Algorithm for Constrained NTF

```
Input: data tensor $\mathcal{X} \in \mathbb{R}^{N_1\times N_2\times\cdots\times N_M}$, the number of
       basis functions $k$, $\alpha$, $\theta$, $q$, maximum iteration steps
       $T$, error threshold $\epsilon$.
Output: the factorization components matrix $\boldsymbol{A}^{(d)}$, $(d = 1,\ldots,M)$
Initialization: set $\boldsymbol{A}^{(d)} \geqslant 0$, $(d = 1,\ldots,M)$ randomly.
Step 1. Repeat until convergence {
Step 2.     For $d = 1$ to $M$ {
Step 3.         Iterate over every entries of $\boldsymbol{A}^{(d)}$ until convergence
                Calculate $\boldsymbol{A}^{(d)}$ using update rules (18) and (19)
                  $\boldsymbol{A}_{ij}^{(d)} \leftarrow \frac{\boldsymbol{A}_{ij}^{(d)}}{\sum_i \boldsymbol{A}_{ij}^{(d)}}$
                  $\boldsymbol{Z}^{(d)} = [\boldsymbol{A}^{(d-1)} \odot \cdots \odot \boldsymbol{A}^{(1)} \odot \boldsymbol{A}^{(M)} \odot \cdots \odot \boldsymbol{A}^{(d+1)}]^{\mathrm{T}}$
            }
Step 4.     Calculate the LS or KL update error $e$ by (14) and (15)
Step 5.     Check convergence: the factorization stage of
            cNTF converges if iteration number exceeds $T$
            or update error $e < \epsilon$
        }
```

For cNTF algorithm at LS case, the major operation in (18) is the matrix product $X_{(d)}Z^{(d)T}S^{T}$ and $SZ^{(d)}Z^{(d)T}S^{T}$, which takes $O(N_dR\prod_{i\neq d}^{M}N_i + N_dR^2)$ and $O(2R^2\prod_{i\neq d}^{M}N_i + R^3)$. When $R \ll \prod_{i\neq d}^{M}N_i$, the time complexity of cNTF at LS case is $O(TR^2(2RM + \sum_{d=1}^{M}N_d)\prod_{i=1}^{M}N_i)$, where T is the maximum iteration number. Similarly, the time complexity of cNTF algorithm at KL case in (19) is same as the LS case. Table 3 gives the time complexity comparison between cNTF, PARAFAC and Tucker algorithm. Here $J_k|_{k=1}^{M}$ is the dimension of core tensor for Tucker algorithm. The comparison results in Table 3 indicate cNTF algorithm is time-consuming in some way. From our practical experimental results, cNTF algorithm is more efficient and provides better recognition performance than other algorithms.

Table 3. Computational Cost Comparison Between cNTF, PARAFAC, Tucker Algorithms

Algorithm	Time Complexity
cNTF	$O(TR^2(2RM + \sum_{d=1}^{M}N_d)\prod_{i=1}^{M}N_i)$
PARAFAC	$O(TMR\prod_{i=1}^{M}N_i)$
Tucker	$O(T(M-1)\sum_{k=1}^{M}J_k\prod_{i=1}^{M}N_i)$

3 Cortical Representation Based on Tensor Structure

In this section, we employ multi-resolution spectro-temporal modulation filters to model the primary auditory cortical representation based on tensor structure. The power spectrum is represented in a multi-linear feature space by a population of cortical cells. cNTF is applied to learn the basis functions of cortical representation. Our proposed method is to explore some intrinsic attributes and mechanism for aiding the design of robust speaker recognition system.

3.1 Cortical Response in Primary Auditory Cortex

In the early stages of auditory processing, speech signals undergo a complex series of transformations and are converted into a 2-dimensional pattern that we call auditory spectrum. The peripheral auditory system decomposes the speech into topographically organized array of channels that are tuned to different center frequencies (CF's). These CF's are similar to logarithmic frequencies and create the tonotopic axis of the auditory system[34]. The auditory spectrum[35-36] represents the neural activity distributed along the tonotopic axis.

The cortical stage concentrates on the neural response of higher central auditory system such as A1 and estimates the spectral-temporal modulation content of an auditory spectrum[35]. The spectrum is decomposed into more elaborate representations and separated into the cues and features associated with different sound percepts such as pitch and timbre.

The neurons in A1 are well organized and have systematic response selectivity to various stimulus features. Every cell is tuned to a specific range of center frequencies and intensities called the response area of the cells[34]. Similar to the topographic characteristic of the receptive fields in visual system, the response areas of A1 also have topographical organization which originates from the segregation of neural response selectivity. Furthermore, it has also been observed that many neurons in A1 are selective to the scale and rate of a frequency modulation (FM) tone[37]. As described in [14], the response areas in A1 are organized along three axes: tonotopic axis, symmetry axis and scale axis. The tonotopic axis gives the change of CF's, the symmetry axis describes asymmetries of response area and the scale axis reflects the bandwidth of each response area along the tonotopic frequency axis.

The above observation suggests that it is reasonable to model the output cortical representation as a higher order tensor structure with three independent modes: the center frequency x, the scale (spectral bandwidth) s and the phase (local symmetry) ϕ. Fig.1 is a sketch of the cortical model for A1. The cuboid in Fig.1 can be seen as arranged with neurons (black dots). Each dot represents a neuron which has its own (x, s, ϕ) coordinates. One example shows the cortical response in the case $\phi = 0$. The curve that the arrow points to reflects the response magnitude of corresponding neuron. The neural response areas have a centered excitatory band that is symmetrically flanked by inhibitory side bands. While as ϕ increases above 0, the response areas become more asymmetric with stronger inhibitory sidebands above CF in one direction and below the CF in the opposite direction.

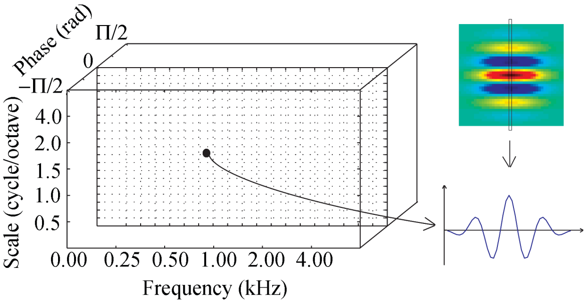


Fig.1. Cortical model of primary auditory cortex.

3.2 Gabor Functions

Inspired by recent physiological and psychoacoustic experimental results for auditory system, much insight has been obtained from the measurements of so-called spectro-temporal response field (STRF) of primary auditory cortex (A1) cell. STRF summarizes the way that neuron cell responds to the stimulus. The neuronphysiological evidence[35] indicates that the cells in the auditory cortex are tuned to localized spectro-temporal modulations. The STRF of these cortical cells[38] can be modeled by 2D Gabor functions. The 2D Gabor filterbank-based method[39] transforms spectrogram into local constituent spectro-temporal amplitudes, frequencies, orientations and phases. In this paper, we model the cortical representation based on tensor structure by the 2D-complex Gabor function $g_{u,v}(f, t)$, which is the product of a Gaussian envelope and a complex plane wave,

$$g_{u,v}(f,t) = g_{\bar{\boldsymbol{k}}}(\bar{\boldsymbol{x}}) = \frac{\bar{\boldsymbol{k}}^2}{\sigma^2} \cdot e^{-\frac{\bar{\boldsymbol{k}}^2 \cdot \bar{\boldsymbol{x}}^2}{2\sigma^2}} \cdot [e^{i\boldsymbol{k}^{\bar{}} \cdot \bar{\boldsymbol{x}}} - e^{-\frac{\sigma^2}{2}}], \quad (20)$$

where $\bar{\boldsymbol{x}} = (f, t)$ is a sample of the power spectrum at frequency f and time frame t, $\bar{\boldsymbol{k}}$ is a vector, which determines the direction and scale of Gabor functions $\bar{\boldsymbol{k}} = k_v e^{i\phi}$, where $k_v = 2^{-\frac{v+2}{2}} \cdot \pi$, $\phi = u\frac{\pi}{K}$, u determines the direction of Gabor functions, v determines the scale of Gabor functions and K determines the total number of directions. Fig.2 gives examples of the real part of Gabor functions with four different scales and four different directions.

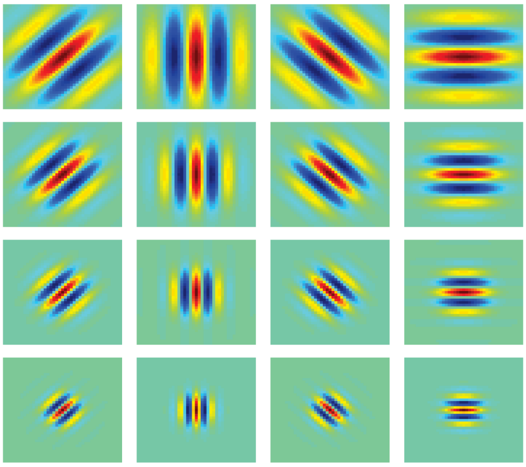


Fig.2. The real part of Gabor functions for four different scales and four different directions.

3.3 Cortical Representation

To explore neural representation in the auditory cortex, we transform speech signals into a form of 4-order tensor. In a given time window, the power spectrum $X(f,t) \in \mathcal{R}^{N_f \times N_t}$ can be represented as a 4-order tensor $\mathcal{X} \in \mathcal{R}^{N_f \times N_t \times N_u \times N_v}$. From Fig.3 we can see clearly that the cortical representation has different spectral patterns which lie on the neuron response area.

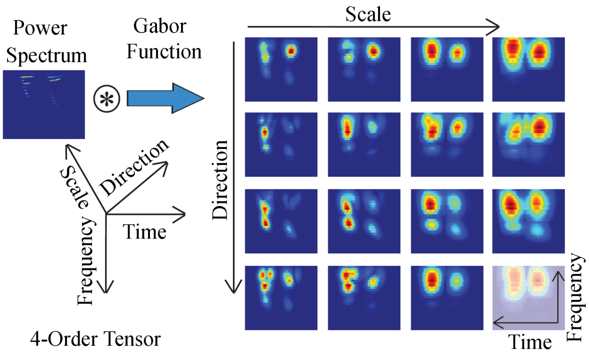


Fig.3. Cortical representation with different parameters. The rows show different directions and the columns show different scales for the power spectrum.

The cortical representation is calculated by convolving the Gabor functions $g_{u,v}(f,t)$ with the power spectrum $X(f,t)$. The result is a 4-order tensor $\mathcal{X} \in \mathcal{R}^{N_f \times N_t \times N_u \times N_v}$, where the first two indices give the time and frequency axes, the third index gives the direction parameters, and the fourth index gives the value of scale. We select the magnitude part of this tensor shown in Fig.3 as our Gabor-based speech feature after the Gaobr filtering. For a Gabor function with fixed scale and direction parameters, the convolution result can be defined as

$$G_{u,v}(f,t) = |X(f,t) \circledast g_{u,v}(f,t)|. \tag{21}$$

The convolution results $G_{u,v}(f,t)$ are spectro-temporal features with different filter characteristics to investigate the multilinear feature space. Here we employ mel-scale filterbanks to map the actual frequency into perceived frequency without losing useful auditory information. The filtered results $G^m_{u,v}(f,t)$ are obtained by a set of critical bands triangular filters which are linear below 1kHz and logarithm above.

3.4 Tensor Analysis and Sparseness Constraint

In order to extract the robust speech feature based on the tensor structure, we transform the Gabor-based multi-resolution representation into multiple interrelated subspaces by cNTF to learn the basis functions $\boldsymbol{A}_l$, $(l = 1, 2, 3, 4)$. Compared with traditional subspace learning methods, the extracted tensor features may characterize the elaborate spectro-temporal patterns of cortical representation and preserve the discriminative information for recognition.

As described in Subsection 3.1, the Gabor-based features can be considered as neuron responses in the primary auditory cortex. Here we employ the sparse localized basis functions $\boldsymbol{A}_1 \in R^{N_f \times d}$ in time-frequency subspace to transform the auditory feature into the sparse feature subspace, where d is the dimension of sparse feature subspace. The sparse feature representation $S_{u,v}$ is obtained from:

$$S_{u,v}(f,t) = \boldsymbol{A}_1^{\mathrm{T}} * G^m_{u,v}(f,t). \tag{22}$$

Fig.4 is an example of basis functions in spectro-temporal domain. From this result we can see that most elements of this project matrix are near to zero, which

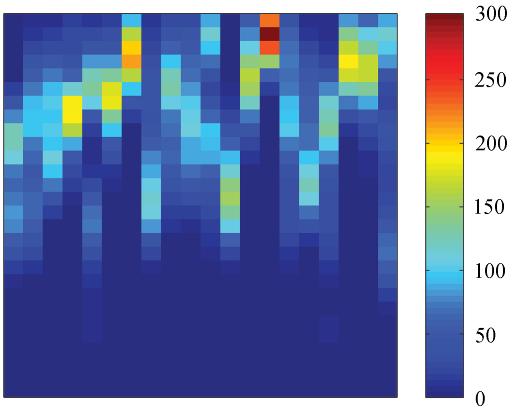


Fig.4. Basis functions of cNTF in spectro-temporal domain.

accords with the sparse constraint of cNTF. The intuitive interpretation why the sparse constraint results in feature robustness is that in sparse coding the energy of signal is only concentrated on a few components, while the energy of additive noise remains uniformly spread on all the components. Noise is reduced for the sparse projection while the useful sparse information is not strongly affected.

Fig.5 shows a comparison between CTCC features and MFCC features for digit *two* before discrete cosine transform (DCT) operation with different signal-to-noise ratio (SNR) levels (clean, 20dB, 10dB). Fig.5(a) presents the standard MFCC features with additive car noise. The degradation of spectral features for MFCC is evident. In Fig.5(b) sparse tensor features with different Gabor function parameters are shown and we can see that the spectral features based on cortical representation after transformation maintain most useful information compared with feature in clean environments.

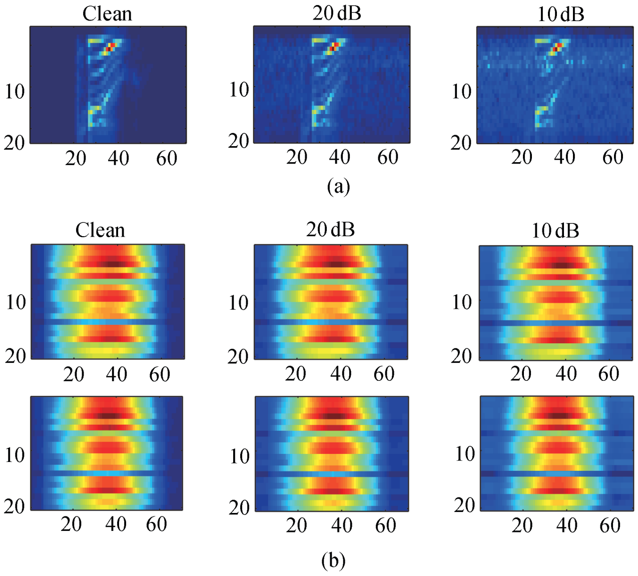


Fig.5. (a) Standard MFCC features. (b) Spare Gabor tensor features with $u = 0$, $\pi/4$ and $v = 5$ before DCT under different SNR conditions (Clean, 20 dB, 10 dB).

The framework of feature extraction is shown in Fig.6. cNTF is employed to learn the basis functions of cortical representation. In a given time window, the cortical representation $\mathcal{X} \in R^{N_f \times N_t \times N_u \times N_v}$ can be calculated by convolving the Gabor functions $g_{u,v}(f,t)$ with the power spectrum $X(f,t) \in R^{N_f \times N_t}$. Then we employ the cortical representation $\mathcal{X}$ as the input tensor data and learn the basis functions $\boldsymbol{A}_l$, $(l = 1, 2, 3, 4)$ by cNTF.

In the GMM model training and testing stage, the robust feature can be extracted by the following steps:

1) employ pre-emphasis on the speech signal and obtain the power spectrum $X(f,t)$ via Fast Fourier Transformation (FFT);

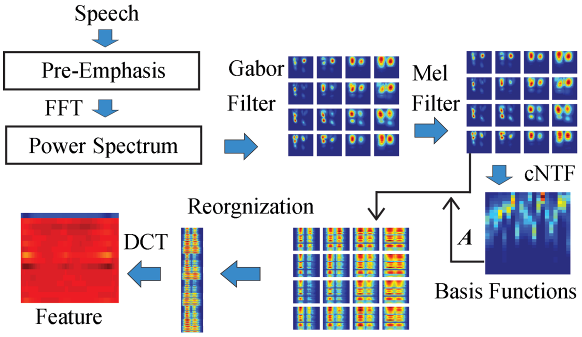


Fig.6. Feature extraction framework based on cortical representation and tensor structure.

2) calculate the cortical representation $\mathcal{X}$ by convolving Gabor functions with the power spectrum;

3) employ the sparse localized basis functions $\boldsymbol{A}_1$ to transform the cortical representation into sparse feature subspace as (22) and obtain the spare tensor features $\mathcal{S}$;

4) reorganize $\mathcal{S}$ into a feature matrix F_G and perform discrete cosine transform (DCT) on spectral feature vectors to reduce the dimensionality and de-correlate the feature components.

4 Experiments and Discussion

In this section we describe the application of CTCC features to the task of speaker recognition in noisy conditions. Moreover, for the comparison purposes, we evaluate the performance of features MFCC, Mel-PCA, Mel-NMF, PARAFAC and Tucker algorithms. Finally, we discuss several issues about our model.

4.1 Experimental Setup

In this paper, Aurora2 speech corpus is used to test the recognition performance, which is designed to evaluate speech recognition algorithms in noisy conditions. In this corpus noise is added to the filtered clean data using the ITU recommendation. The training corpus includes 8440 sentences of digit sequences (about 76 from each of 110 speakers) in clean condition for building recognition model. Three different test datasets (*TestA*, *TestB*, *TestC*) in clean and noise conditions (20 dB, 15 dB, 10 dB, 5 dB, 0 dB, −5 dB) are taken for the recognition.

In our experiment, the sampling rate of speech signal was 8 kHz. To compute the power spectrum, a Hamming window of 25 ms was shifted over an input speech utterance every 10 ms. In each frame, a segmented utterance was converted to its corresponding 256-dimensional FFT-based power spectrum vector. The multi-resolution Gabor-based features were derived from the power spectrum by Gabor functions with 4 different scales and 4 different directions. The output magnitude results were filtered by 40-channel Mel filterbanks to create the tensor representations for tensor factorization. We randomly selected 110 sentences as training data to learn the basis functions using cNTF after the calculation of cortical tensor features.

In order to estimate the speaker model and test the efficiency of our method, we use 5500 sentences (50 sentences each person) as training data which is independent with the data samples for basis functions learning and 1320 clean sentences (12 sentences each speaker) are selected for testing. The testing samples in noise condition are created by former clean test sentences mixed with subway, babble, car noise, exhibition hall in SNR intensities of 20 dB, 15 dB, 10 dB and 5 dB respectively. For the final feature set, 16 cepstral coefficients were extracted and used for speaker modeling. GMM was used to build the recognizer with 64 Gaussian mixtures.

4.2 Experimental Results

For comparison, the performance of MFCC, Mel-NMF, Mel-PCA, PARAFAC and Tucker algorithms with 16-order cepstral coefficients are also tested. We use PCA and NMF to learn the part-based representation in the spectro-temporal domain after Mel filtering, which is similar to [40]. The feature after PCA or NMF projection was further processed into the cepstral domain via discrete cosine transform. PARAFAC and Tucker algorithms are used to learn the basis functions and the feature extraction procedure is same as our framework.

The identification accuracy results obtained by CTCC and baseline system in all testing conditions are summarized in Table 4. We can observe from Table 4 that the performance degradation of CTCC is slower with increasing noise intensity compared with other features. It performs better than other features in the high noise conditions such as 5dB condition noise. Fig.7 describes the identification rate in four noisy conditions averaged over SNRs between 5∼20 dB, and the overall average accuracy across all the conditions. The results suggest that this cortical representation feature is robust against the additive noise, which indicates the potential of the new feature for dealing with a wider variety of noisy conditions.

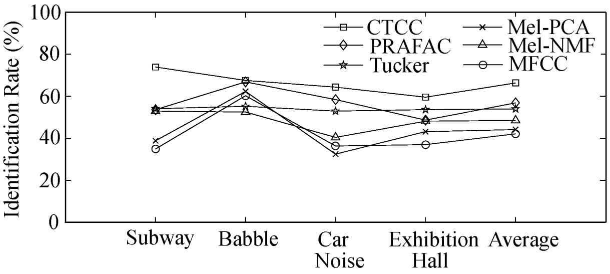


Fig.7. Identification accuracy in four noisy conditions averaged over SNRs between 5∼20dB, and the overall average accuracy across all the conditions, for CTCC and other features using Aurora2 noise testing dataset.

4.3 Discussion

We generalize the common NTF algorithm by the nonsmooth approach and orthogonal constraint. Smoothing matrix $\boldsymbol{S}$ is introduced to control the sparseness of basis function and encoding matrix in each

Table 4. Identification Accuracy in Four Noisy Conditions for Aurora2 Noise Testing Dataset

Noise	SNR (dB)	CTCC (%)	MFCC (%)	PARAFAC (%)	Tucker (%)	NMF (%)	PCA (%)
Subway	5	34.55	2.73	7.27	19.09	15.45	3.64
	10	75.45	16.36	39.09	40.90	40.91	12.73
	15	89.09	44.55	75.45	71.81	67.27	50.91
	20	96.36	76.36	92.73	84.54	88.18	88.18
Babble	5	28.18	16.36	23.64	15.45	23.64	21.82
	10	57.27	51.82	60.00	45.45	41.82	51.82
	15	88.18	79.09	86.36	70.00	61.82	79.09
	20	96.36	93.64	97.27	90.00	82.73	96.36
Car Noise	5	22.73	5.45	14.55	15.45	3.64	2.73
	10	54.55	17.27	52.73	40.00	26.36	10.00
	15	81.82	44.55	75.45	69.09	57.27	38.18
	20	98.18	78.18	90.91	87.27	74.55	79.09
Exhibition Hall	5	23.64	1.82	9.09	13.63	9.09	3.64
	10	47.27	20.00	28.18	36.36	29.09	20.91
	15	75.45	50.00	69.09	72.72	68.18	59.09
	20	91.82	76.36	88.18	91.81	86.36	89.09

mode. For cNTF algorithm, $\boldsymbol{S}$ will control the sparseness of mode matrices $\boldsymbol{A}^{(d)}$. By this sparseness control operator, cNTF can produce more localized and less overlapped representations which enhance the robustness of features and reduce the noise components. The orthogonal constraint $\sum_{p\neq q}[\boldsymbol{A}^{(d)T}\boldsymbol{A}^{(d)}]_{pq}$ makes the basis functions be as orthogonal as possible to ensure redundancy minimization between different basis functions.

Motivated by the cortical representation in the primary auditory cortex, we model the neuron response by the 2D Gabor function with different scales and directions. It is assumed that the Gabor functions are similar to the receptive field profiles in the mammalian cortical simple cells. The cortical stage in central auditory system analyzes the spectrum into more elaborate multiple factor representations. These representations reflect the neuron response for different perception cues and will enhance the robustness of features.

Under the multilinear analysis framework, we employ cNTF to learn the basis functions of cortical representation. The basis functions with sparse constraint fit the statistical characteristic of the clean speech data. The sparse assumption makes the energy of signal concentrate on a few components. After transformation, the components that accord with former statistical characteristic will be preserved, while the noise components with different distributions will be suppressed.

Traditional spectral analysis methods such as MFCC provide an approximation of frequency integration of the auditory system in the 2D spectro-temporal feature space. Actually, as described in Subsection 3.1, the feature space is a higher order tensor space with four independent modes. Therefore we believe that the spectro-temporal patterns of speech in different feature subspaces provide more robust information for speech feature analysis. This model can be considered as the paralleled spectral analysis within the context of auditory framework.

5 Conclusion

In this paper, we derive a new feature extraction method called cNTF for learning basis functions in multiple interrelated subspaces. Sparse constraint on cNTF enhances energy concentration of speech signal, keeping useful features during noise reduction. Furthermore the orthogonal constraint ensures redundancy minimization between different basis functions. A novel robust feature extraction framework for speech is proposed based on integration cortical representation and tensor factorization, inspired by the multilinear representation of central auditory system. Our approach is primarily data-driven and effectively extracts robust feature of speech called CTCC which is invariant to the types and interference of noise with different intensities. The cortical sparse representation is extracted after the multi-related subspace projection, preserving the discriminative and robust information of different speakers. Experimental results demonstrate that the CTCC feature improves the noisy robustness and recognition accuracy compared with baseline systems.

References

[1] Rabiner L R, Juang B. Fundamentals on Speech Recognition. New Jersey: Prentice Hall, 1996.

[2] Hermansky H. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 1990, 87(4): 1738-1752.

[3] Reynolds D A, Rose R C. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. Speech and Audio Processing*, , 1995, 3(1): 72-83.

[4] Hermansky H, Morgan N. RASTA processing of speech. *IEEE Trans. Speech and Audio Processing*, 1994, 2(4): 578-589.

[5] Reynolds D A. Experimental evaluation of features for robust speaker identification. *IEEE Trans. Speech and Audio Processing*, 1994, 2(4): 639-643.

[6] Mammone R, Zhang X, Ramachandran R P. Robust speaker recognition: A feature-based approach. *IEEE Signal Process. Mag*, 1996, 13(5): 58-71.

[7] Van Vuuren S. Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch. In *Proc. ICSLP 1996*, Oct. 3-6, 1996, Vol.3, pp.1788-1791.

[8] Berouti M, Schwartz R, Makhoul J, Beranek B, Newman I, Cambridge M A. Enhancement of speech corrupted by acoustic noise. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP 1979*), Washington DC, USA, April 2-4, 1979, Vol.4, pp.208-211.

[9] Wu M Y, Wang D L. A two-stage algorithm for one-microphone reverberant speech enhancement. *IEEE Transactions on Speech and Audio Processing*, 2006, 14(3): 774-784.

[10] Hu Y, Loizou P C. A perceptually motivated subspace approach for speech enhancement. In *Proc. the Seventh International Conference on Spoken Language Processing*, Denver, USA, Sept. 16-20, 2002.

[11] Hermus K, Wambacq P, Van Hamme H. A review of signal subspace speech enhancement and its application to noise robust speech recognition. *EURASIP Journal on Applied Signal Processing*, 2007, (1): 195-209.

[12] Mami Y, Charlet D. Speaker identification by anchor models with PCA/LDA post-processing. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP 2003*), Hong Kong, China, April 6-10, 2003, pp.180-183.

[13] Wilson K W, Raj B, Smaragdis P, Divakaran A. Speech denoising using nonnegative matrix factorization with priors. *IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP 2008*), Las Vegas, USA, March 30-April 4, 2008, pp.4029-4032.

[14] Wang K, Shamma S A. Spectral shape analysis in the central auditory system. *IEEE Transactions on Speech and Audio Processing*, 1995, 3(5): 382-395.

[15] Yang X, Wang K, Shamma S A. Auditory representation of acoustic signals. *IEEE Trans. Information Theory*, 1992, 38(2): 824-839.

[16] Mesgarani N, Slaney M, Shamma S A. Discrimination of speech from nonspeech based on multiscale spectro-temporal Modulations. *IEEE Trans. Audio, Speech, and Language Processing*, 2006, 14(3): 920-930.

[17] Woojay J, Juang B H. Speech analysis in a model of the central auditory system. *IEEE Trans. Audio, Speech, and Language Processing*, 2008, 15(6): 1802-1817.

[18] Carroll J D, Chang J J. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 1970, 35(3): 283-319.

[19] Harshman R A. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. UCLA Working Papers in Phonetics, 1970, 16: 1-84.

[20] Bro R. PARAFAC: Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 1997, 38(2): 149-171.

[21] Kroonenberg P M, de Leeuw J. Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, 1980, 45(1): 69-97.

[22] Lathauwer L D. Signal processing based on multilinear algebra [Ph.D. Dissertation]. Katholike Universiteit Leuven, 1997.

[23] Lathauwer L D, Moor B D, Vandewalle J. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 2000, 21(4): 1253-1278.

[24] Welling M, Weber M. Positive tensor factorization. *Pattern Recognition Letters*, 2001, 22(12): 1255-1261.

[25] Shashua A, Hazan T. Non-negative tensor factorization with applications to statistics and computer vision. In *Proc. IEEE International Conference on the International Conference on Machine Learning* (*ICML*), Bonn, Germany, Aug. 7-11, 2005, pp.792-799.

[26] Cichocki A, Zdunek R, Choi S, Plemmons R, Amari S. Non-negative tensor factorization using alpha and beta divergences. In *Proc. Acoustics, Speech and Signal Processing*, Honolulu, USA, April 15-20, 2007, Vol.3, pp.1393-1396.

[27] Vasilescu M A O, Terzopoulos D. Multilinear independent components analysis. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, San Diego, USA, Jan. 20-26, 2005, Vol.1, pp.547-553.

[28] Tao D C, Li X L, Wu X D, Maybank S J. General tensor discriminant analysis and Gabor feature for gait recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29(10): 1700-1715.

[29] Tao D C, Li X L, Wu X D, Maybank S J. Tensor rank one discriminant analysis — A convergent method for discriminative multilinear subspace selection. *Neurocomputing*, 2008, 71(10-12): 1866-1882.

[30] Stegeman A, Sidiropoulos N D. On Kruskal's uniqueness condition for the Candecomp/Parafac decomposition. *Linear Algebra and Its Applications*, 2007, 420(2/3): 540-552.

[31] Comon, P. Mathematics in Signal Processing V. Oxford University Press, USA, 2002.

[32] Lee D D, Seung H S. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 2001, 13: 556-562.

[33] Pascual-Montano A, Carazo J M, Kochi K, Lehmann D, Pascual-Marqui R D. Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006, 28(3): 403-415.

[34] Merzenich M M, Knight P L, Roth G L. Representation of cochea within the primary auditory cortex in cat. *Journal of Neurophysiology.* 1975, 38(2): 231-249.

[35] Chi T, Ru P, Shamma S A. Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 2005, 118(2): 887-906.

[36] Wang K, Shamma S A. Self-normalization and noise-robustness in early auditory representations. *IEEE Transactions on Speech and Audio Processing*, 1994, 2(3): 421-435.

[37] Mendelson J R, Cynader M S. Sensitivity of cat primary auditory cortex (AI) neurons to the direction and rate of frequency modulation. *Brain Research*, 1985, 327(1/2): 331-335.

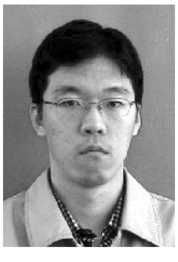
[38] Qiu A, Schreiner C E, Escabi M A. Gabor analysis of auditory midbrain receptive fields: Spectro-temporal and binaural composition. *Journal of Neurophysiology*, 2003, 90(1): 456-476.

[39] Ezzat T, Bouvrie J, Poggio T. Max-Gabor analysis and synthesis of spectrograms. In *Proc. Ninth International Conference on Spoken Language Processing* (*ICASLP 2006*), Pittsburg, USA, Sept. 17-21, 2006.

[40] Cho Y C, Choi S. Nonnegative features of spectro-temporal sounds for classification. *Pattern Recognition Letters*, 2005, 26(9): 1327-1336.

Qiang Wu received the B.E. degree in computer science from Liaocheng University, China in 2002. He received the M.E. degree in computer science from Shandong University (SJTU), China in 2005. He is currently a Ph.D. candidate of Department of Computer Science and Engineering, SJTU, Shanghai, China. His research interests include speech signal processing, brain-computer interface, machine learning, neuroscience.

Li-Qing Zhang received the Ph.D. degree from Zhongshan University, Guangzhou, China, in 1988. He was promoted to full professor position in 1995 at South China University of Technology. He worked as a research scientist in RIKEN Brain Science Institute, Japan from 1997 to 2002. He is now a professor with Department of Computer Science and Engineering, SJTU. His current research interests cover computational theory for cortical networks, brain signal processing and brain-computer interface, perception and cognition computing model, statistical learning and inference. He has published more than 160 papers in international journals and conferences.



Guang-Chuan Shi received the B.E. degree in computer science and engineering from SJTU, in 2008. He is currently an M.E. candidate of SJTU. His research interests include signal processing and perceptual computation.