Robust Multifactor Speech Feature Extraction Based on Gabor Analysis

Qiang Wu, Liqing Zhang, and Guangchuan Shi

***Abstract*—The performance of speech recognition systems relies on the consistency and adaptation of the speech feature in complex conditions during the training and testing stages. Traditional systems usually perform poorly under adverse noisy conditions and are not applicable to most real world problems. In this paper, we investigate the speech feature extraction problem in a noisy environment and propose a novel approach based on Gabor filtering and tensor factorization. Recent physiological and psychoacoustic experimental results suggest that the localized spectro-temporal features are essential for auditory perception. To explore this property, we represent the speech signal by using a general higher order tensor and employ two-dimensional Gabor functions with different scales and directions to analyze the localized patches of the power spectrogram. Then the Nonnegative Tensor PCA with sparse constraints is proposed to learn the projection matrices from multiple interrelated feature subspaces. The objective of the sparse constraints is to preserve the statistical characteristic of clean speech data by finding projection matrices of speech subspaces and reduce the noise components which have distributions different from those of clean speech. A multifactor analysis method is proposed to extract robust sparse features by processing the data samples in tensor structure. The simulation results indicate that our proposed method is able to improve the speech recognition performance, especially in noisy environments, compared with the traditional speech feature extraction methods.**

***Index Terms*—Acoustic noise, auditory perception, feature extraction, Gabor filtering, speech recognition, tensor factorization.**

I. Introduction

In SPEECH recognition systems, feature extraction and recognition are two important modules. The primary objective of feature extraction is to find robust and discriminative features in the acoustic data. The recognition module uses the speech features and the acoustic models to decode the speech input and produces text results with high accuracy. A number of speech feature extraction methods have been proposed, such as linear predictive cepstral coefficients (LPCCs) [1], mel-frequency cepstral coefficients (MFCCs) and perceptual linear predictive coefficients (PLPs) [2].

Manuscript received January 29, 2010; revised May 09, 2010; accepted August 05, 2010. Date of publication August 26, 2010; date of current version March 30, 2011. This work was supported in part by the Science and Technology Commission of Shanghai Municipality under Grant 08511501701 and in part the National Natural Science Foundation of China under Grant 60775007, 90920014. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Haizhou Li.

The authors are with the MOE-Microsoft Key Laboratory for Intelligent Computing and Intelligent Systems, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: johnnywu@sjtu.edu.cn; johnstonwu@gmail.com; lqzhang@sjtu.edu.cn; sgc1984@sjtu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2070495

Several available speech recognition systems are able to achieve acceptable accuracy for clean speech, while the recognition performance is degraded dramatically in noisy environments. Performance degradation is attributed to the inevitable mismatch of speech features between training and testing conditions. Several methods have been proposed to reduce the effect of mismatch. Feature compensation techniques such as CMN [3] and RASTA [4] have been developed for robust speech recognition. Given *a priori* knowledge of the noise spectrum, the effectiveness of subspace-based filtering and spectral subtraction techniques has been shown in [5]–[7]. Temporal filtering approaches [8], [9] based on specific optimization techniques have been proved to be capable of enhancing the discrimination and robustness of speech features in speech recognition. Temporal structure normalization (TSN) [10] which normalizes the temporal statistics of the speech features aims to reduce the noise disorders.

Recently, computational auditory neural models and sparse coding have attracted much attention in the societies of neuroscience and speech processing. The Gabor STRF model [11] was proposed to fit the auditory nucleus of inferior colliculus by using spectral and temporal Gabor functions. Kleinschmidt [12] provided a brief overview on the work related to the localized spectro-temporal features. In [13], a 2-D spectro-temporal Gabor filterbank based on 2-D fast Fourier transform (FFT) was developed to decompose the spectrogram patch into underlying dominant spectro-temporal components. In [14], multiscale spectro-temporal modulation features based on auditory cortex model were proposed and HOSVD was applied to perform multilinear dimensionality reduction for investigating content-based audio classification problem. Jeon [15] proposed a computational central auditory system model and interpreted various feature selection methods.

Multifactor analysis provides a potential approach for generating robust features that can discriminate speech entities better. As a powerful data modeling tool for pattern recognition, multilinear algebra of higher order tensors has been proposed as a potent mathematical framework to manipulate the multiple factors underlying the observations. Tensor decomposition methods include CANDECOMP/PARAFAC model [16]–[18], the Tucker Model [19], [20], and Nonnegative Tensor Factorization (NTF) [21], [22]. Tensor discriminant analysis methods [23], [24] were proposed to deal with the small sample size problem that occurs in conventional discriminant learning.

In this paper, we propose a robust speech feature extraction method based on Gabor analysis and tensor factorization. 2-D Gabor functions are used to extract the spectro-temporal information, which employ multi-resolution wavelet over scales

and directions to analyze the power spectrogram of speech. A new tensor analysis approach called NTPCA is developed for multifactor analysis of speech by maximizing the covariance of data samples on the tensor structure. The Gabor tensor feature extracted by NTPCA can be processed further into a representation called Gabor tensor cepstral coefficients (GTCCs) by discrete cosine transform (DCT). Finally hidden Markov models (HMMs) are used to construct a speech recognizer with GTCC features. The advantages of our method include the following. 1) The features based on multi-resolution spectro-temporal modulation with different scales and directions are biologically plausible, which simulate the auditory cortical representation. The speech signal can be represented in the framework of a higher order tensor so that both spectral and temporal structures can be explored simultaneously within one model. 2) A supervised learning procedure is proposed to find the projection matrices of multi-related feature subspaces preserving the individual, spectro-temporal information in the tensor structure. The maximum variance criterion avails to remove the noise component as useless information in the minor subspace. 3) The sparse constraints make the energy of speech signal concentrate on a few components and the statistical characteristics of clean speech data are preserved in the projection matrices. Therefore, the projected components with consistent statistical characteristic will be reserved, while the noise components with different distributions will be suppressed. To validate the performance of the proposed method, comparisons with commonly used feature extraction methods are given. Experimental results show GTCC features are good speech representations, reserving robust features in noisy environments.

The remainder of this paper is organized as follows. In Section II, a tensor factorization algorithm NTPCA is presented for feature extraction. Section III describes the Gabor tensor feature extraction framework for robust speech recognition. Section IV presents the experimental results of speech recognition in the noise-free and noisy environments. Finally, Section V provides a summary and conclusions.

II. Nonnegative Tensor Principal Component Analysis

In this section, we introduce nonnegative tensor principal component analysis (NTPCA) [28] as an extension of nonnegative sparse principle component analysis to preserve the intrinsic structure of high-order tensor data and avoid the lose of spatial information. The time complexity and convergence analysis of NTPCA are also provided.

A. Fundamentals of Multilinear Algebra

For the integrity of this paper, we briefly introduce relevant definitions of multilinear algebra, which is fundamental to this paper. More details can be found in [19], [23], [25]. A tensor is a multidimensional array that is an element of the tensor product of N vector spaces, each of which has its own coordinate system. The order of a tensor $\mathbf{X} \in R^{N_1 \times N_2 \times \cdots \times N_M}$ is M, that is the number of factors, also known as ways or modes. An element of $\mathbf{X}$ is denoted by $\mathbf{X}_{n_1,n_2,\ldots,n_M}$, where $1 \le n_i \le N_i$ and $1 \le i \le M$.

Definition 1 (Tensor Product): The tensor product $\mathbf{X} \otimes \mathbf{Y}$ of a tensor $\mathbf{X} \in R^{N_1 \times \cdots \times N_M}$ and another tensor $\mathbf{Y} \in R^{N'_1 \times N'_2 \times \cdots \times N'_M}$ is a tensor defined by

$$(\mathbf{X} \otimes \mathbf{Y})_{n_1,n_2,\ldots,n_M,n'_1,n'_2,\ldots,n'_M} = \mathbf{X}_{n_1,n_2,\ldots,n_M} \mathbf{Y}_{n'_1,n'_2,\ldots,n'_M} \tag{1}$$

for all index values.

Definition 2 (Matrix Unfolding): The matrix unfolding or mode-d matricizing of an M-order tensor $\mathbf{X} \in R^{N_1 \times N_2 \cdots \times N_M}$ is the set of vectors in R^{N_d} obtained by keeping the dth index fixed and varying the other indices. The matrix unfolding or mode-d matricizing of an M-order tensor is a matrix $\mathbf{X}_{(d)} \in R^{N_d \times \bar{N}_d}$, where $\bar{N}_d = \prod_{i \ne d} N_i$. We denote the mode-$d$ matricizing of $\mathbf{X}$ as $mat_d(\mathbf{X})$ or $\mathbf{X}_{(d)}$.

Definition 3 (Tensor Contraction): In this paper, the contraction of a tensor is conducted on all indices except the ith index on the tensor product of $\mathbf{X} \in R^{N_1 \times N_2 \times \cdots \times N_M}$ and $\mathbf{Y} \in R^{N_1 \times N_2 \times \cdots \times N_M}$. We denote the contraction result as

$$\begin{aligned}
&([\mathbf{X} \otimes \mathbf{Y}; (\bar{i})(\bar{i})])_{j,k} \\
&= [\mathbf{X} \otimes \mathbf{Y}; (1:i-1, i+1:M)(1:i-1, i+1:M)] \\
&= \sum_{n_1=1}^{N_1} \cdots \sum_{n_{i-1}=1}^{N_{i-1}} \sum_{n_{i+1}=1}^{N_{i+1}} \\
&\quad \cdots \sum_{n_M=1}^{N_M} \mathbf{X}_{n_1,\ldots,n_{i-1},j,n_{i+1},\ldots,n_M} \mathbf{Y}_{n_1,\ldots,n_{i-1},k,n_{i+1},\ldots,n_M}
\end{aligned} \tag{2}$$

where $j,k = 1,\ldots,N_i, [\mathbf{X} \otimes \mathbf{Y}; (\bar{i})(\bar{i})] \in R^{N_i \times N_i}$. Equivalent expression of (2) can be written as

$$[\mathbf{X} \otimes \mathbf{Y}; (\bar{i})(\bar{i})] = mat_i(\mathbf{X}) mat_i^T(\mathbf{Y}) = \mathbf{X}_{(i)} \mathbf{Y}_{(i)}^T \tag{3}$$

Definition 4 (Mode-d Matrix Product): The mode-d matrix product defines multiplication of a tensor with a matrix in mode d. Let $\mathbf{X} \in R^{N_1 \times \cdots \times N_M}$ and $A \in R^{J \times N_d}$

$$\begin{aligned}
&(\mathbf{X} \times_d A)_{n_1,\ldots,n_{d-1},j,n_{d+1}\ldots,n_M} \\
&= \sum_{n_d} (\mathbf{X}_{n_1,\ldots,n_d,\ldots,n_M} A_{j,n_d}) \\
&= [\mathbf{X} \otimes A; (d)(2)]_{n_1,\ldots,n_{d-1},j,n_{d+1}\ldots,n_M}.
\end{aligned} \tag{4}$$

We simplify the notation of mode-d matrix product as

$$\mathbf{X} \times_1 A_1 \times_2 A_2 \times \cdots \times_M A_M = \mathbf{X} \prod_{i=1}^{M} \times_i A_i \tag{5}$$

and

$$\begin{aligned}
&\mathbf{X} \times_1 A_1 \times \cdots \times_{i-1} A_{i-1} \times_{i+1} A_{i+1} \times \cdots \times_M A_M \\
&\qquad = \mathbf{X} \prod_{k=1,k\ne i}^{M} \times_k A_k = \mathbf{X} \overline{\times}_i A_i
\end{aligned} \tag{6}$$

where $A_i \in R^{J_i \times N_i}$.

Definition 5 (Frobenius Norm): The Frobenius norm of a tensor $\mathbf{X} \in R^{N_1 \times N_2 \times \cdots \times N_M}$ is given by

$$\begin{aligned}
\|\mathbf{X}\|_F &= \sqrt{[\mathbf{X} \otimes \mathbf{X}; (1:M)(1:M)]} \\
&= \sqrt{\sum_{n_1=1}^{N_1} \cdots \sum_{n_M=1}^{N_M} \mathbf{X}^2_{n_1,\ldots,n_M}}.
\end{aligned} \tag{7}$$

Obviously the mode-d matricizing of tensor $\mathbf{X}$ has the same Frobenius norm as tensor $\mathbf{X}$, that is $\|mat_d(\mathbf{X})\|_F = \|\mathbf{X}\|_F$.

B. Nonnegative Sparse Principal Component Analysis

Principal component analysis (PCA) is a widely used dimensionality reduction technique in data analysis. In [27], nonnegative sparse principal component analysis (NSPCA) was proposed by incorporating both nonnegativity and sparseness into PCA with maintaining the maximal variance property.

The problem of NSPCA can be formulated as follows: given a set of centered data points $X \in R^{d\times n}$, we are to find a number of principal vectors $\{u_i \in R^d\}|_{i=1}^k$, such that those vectors maximize the following objective function:

$$\bar{U} = \arg\max_U \frac{1}{2}\|U^T X\|_F^2 - \frac{\alpha}{4}\|I - U^T U\|_F^2 - \beta \mathbf{1}^T U \mathbf{1} \quad \text{s.t.} \quad U \geq 0 \quad (8)$$

where $X \in R^{d\times n}$ and $U \in R^{d\times k}$, $\|A\|_F^2$ is the square Frobenius norm, the second term in (8) is to relax orthogonal constraints in traditional PCA, $\alpha > 0$ controls the additional orthogonality required, the third term in (8) is the sparse constraint, $\mathbf{1}$ is a column vector with all elements equal to one, and $\beta \geq 0$ is a parameter for controlling sparseness degree.

The sparse constraints can decrease the density of projection matrices, i.e., reduce the average number of nonzero elements per principal vector. To minimize the number of nonzero elements of a principal vector, we can impose the L_0 norm constraint on the principal vector. Generally, L_0 constraint will bring in computational complexity in finding the optimal solution. Therefore, we replace L_0 constraint with L_1 norm constraint, i.e., $\|U\|_{L_1} = \sum_{i=1}^d \sum_{j=1}^k |u_{ij}|$. Since U is nonnegative we can use the sparseness term: $\|U\|_{L_1} = \mathbf{1}^T U \mathbf{1}$.

C. Nonnegative Tensor Principal Component Analysis

Similar to NSPCA, we denote the ith centered training sample (tensor) as an r-order tensor $\mathbf{X}_i \in R^{N_1\times\cdots\times N_r}$, where $i = 1,\ldots,n$ and $U_l \in R^{N_l\times N_l^*}$, $(l = 1,2,\ldots,r)$ denotes the lth mode projection matrix to be found in training procedure. Denote

$$A_l = \sum_{i=1}^n [mat_l(\mathbf{X}_i \overline{\times}_l U_l^T) mat_l^T(\mathbf{X}_i \overline{\times}_l U_l^T)]. \quad (9)$$

The problem of NTPCA is to find lth projection matrix U_l which maximizes the following optimization problem:

$$\bar{U}_l|_{l=1}^r = \arg\max_{U_l|_{l=1}^r \geq 0} \frac{1}{2}\|U_l^T B_l\|_F^2 - \frac{\alpha_l}{4}\|I - U_l^T U_l\|_F^2 - \beta_l \mathbf{1}^T U_l \mathbf{1} + C_l \quad (10)$$

where B_l is defined by $A_l = B_l B_l^T$, $C_l = -(\alpha_l/4)\sum_{k=1,k\neq l}^r \|I - U_k^T U_k\|_F^2 - \beta_l \sum_{k=1,k\neq l}^r \mathbf{1}^T U_k \mathbf{1}$. Because of the sparse constraints imposed on projection matrices, minimizing $\mathbf{1}^T U_l \mathbf{1}$ will cause some elements of projection matrices U_l to be exactly zero. The detailed derivation of problem foundation (10) is given in Appendix A.

For fixed α_l and β_l, the optimization problem (10) is an NP-hard problem [26], [27]. Then we decompose the optimization problem (10) into an iterative optimization of some variables by fixing the rest of the variables invariant and find the local optimal solutions. To find lth mode projection matrix U_l, we first fix the projection matrices of other modes and solve the optimization problem (10) by iterative procedures. The suboptimization function of u_{lpq} (the qth row of the u_p column vector with index l) is defined as follows:

$$f(u_{lpq}) = -\frac{\alpha_l}{4} u_{lpq}^4 + \frac{c_2}{2} u_{lpq}^2 + c_1 u_{lpq} + \text{const} \quad (11)$$

where

$$c_1 = -\alpha_l \cdot \sum_{i=1,i\neq p}^{N_l^*} \sum_{j=1,j\neq q}^{N_l} u_{lpj} u_{lij} u_{liq} - \beta_l + \sum_{i=1,i\neq q}^{N_l} a_{lqi} u_{lpi}$$

$c_2 = a_{lqq} + \alpha_l - \alpha_l \cdot \sum_{i=1,i\neq q}^{N_l} u_{lpi}^2 - \alpha_l \cdot \sum_{i=1,i\neq p}^{N_l^*} u_{liq}^2$, $const$ is a term independent of u_{lpq} and a_{lij} is the element of A_l. We can calculate the derivative $\partial f(u_{lpq})/\partial u_{lpq}$ with respect to u_{lpq} and set it to zero to obtain the nonnegative roots and zero as the nonnegative global maximum of $f(u_{lpq})$.

Algorithm 1: Algorithm of NTPCA

Data: Training data $\mathbf{X}_i \in R^{N_1\times N_2\times\cdots N_r}$, $i = 1,\ldots,n$, dimensionality of output tensors $N_j^*|_{j=1}^r$, α_l, β_l, $(l = 1,\ldots,r)$, maximum iterations T, error threshold ε.

Result: The projection matrix $U_l \geq 0$ $(l = 1,\ldots,r)$, the output tensors $\mathbf{Y}_j$.

1) Initialization: Set $U_l^{(0)} \geq 0$ $(l = 1,\ldots,r)$ randomly, iteration index $t = 1$;
2) **repeat**
3) **for** $l \leftarrow 1$ to r **do**
4) Calculate $\mathbf{A}_l^{(t-1)}$;
5) Iterate over every entries of $U_l^{(t)}$
6) -Set the value of u_{lpq} to the global nonnegative maximizer of **(11)**.
7) **until** $t > T$ or update error $e < \varepsilon$;
8) $\mathbf{Y}_j = \mathbf{X}_j \prod_{l=1}^r \times_l U_l^T$

The optimization problem (10) can be solved by the alternating projection optimization procedure of the Nonnegative Tensor PCA (NTPCA) (See Algorithm 1). The key steps in the alternating projection procedure are Steps 4–6, which involve finding the lth projection matrix $U_l^{(t)}$ in the tth iteration by using $U_k^{(t-1)}|_{1\leq k\leq r}^{k\neq l}$ found in the $(t-1)$th iteration. In Step 4 we calculate the covariance matrix $A_l^{(t-1)}$ with given $U_k^{(t-1)}|_{1\leq k\leq r}^{k\neq l}$ in the $(t-1)$th iteration. Then $U_l^{(t)}$ is updated by maximizing (11). By iterating Steps 4–6 in Algorithm 1, we can obtain the projection matrices $U_l|_{l=1}^r$ for different modes.

D. Time Complexity Analysis

The time complexity of NSPCA is $O(\hat{T}\prod_{i=1}^r N_i N^{*2})$ when the sample $\mathbf{X}$ belong to $R^{N_1\times N_2\times\cdots N_r}$, where $\hat{T}$ is the maximal iteration number of NSPCA, N^* is the number of selected features. The time complexity of NTPCA based on the alternating projection methods is $O(T\sum_{i=1}^r N_i N_i^{*2})$, where T is the number of iterations to make the optimization procedure NTPCA converge, N_i is the dimension of sample in mode i,

N_i^* is the number of selected features in mode i. The space complexity of the alternating projection optimization of NTPCA is $O(\sum_{i=1}^{r} N_i^2)$.

The tensor representation can reduce the number of parameters for modeling the data. For common subspace methods, a multifactor data structure with sample $\mathbf{X}_i \in R^{N_1 \times N_2 \times N_3}$ is a vector in $R^{N_1 \cdot N_2 \cdot N_3}$. Thus, we need to estimate the projection matrix U in $R^{N_1 N_2 N_3 \times N^*}$ for NSPCA, while we only need to estimate the projection matrices $U_i \in R^{Ni \times N_i^*}$, i =1, 2, 3 in NTPCA. The estimation procedure for each U_i is implemented independently, which makes the number of parameters in NTPCA less than that of NSPCA.

E. Convergence Analysis

Similar to the convergence analysis in [23], [24], the convergence of NTPCA is also guaranteed during the alternating optimization procedure. We define a continuous function $f : S_1 \times S_2 \times \cdots \times S_r = \prod_{l=1}^{r} S_l \rightarrow R^+$

$$\begin{aligned} & f(U_l|_{l=1}^r) \\ & \doteq \frac{1}{2} \\ & \cdot \sum_{i=1}^{n} \left[\left(\mathbf{X}_i \prod_{k=1}^{r} \times_k U_k^T \right) \otimes \left(\mathbf{X}_i \prod_{k=1}^{r} \times_k U_k^T \right); (1:r)(1:r) \right] \\ & - \frac{\alpha_l}{4} \sum_{k=1}^{r} ||I - U_k^T U_k||_F^2 - \beta_l \sum_{k=1}^{r} \mathbf{1}^T U_k \mathbf{1} \end{aligned} \tag{12}$$

where $U_l \in S_l$ with the constraint $U_l \geq 0$, S_l is the set, which includes all possible U_l. As described in (24), the optimization problem of NTPCA can be decomposed into r subproblems. With the definition (12), f has r different mappings, as shown in (13) at the bottom of the page, where $f_l(U_l) = f(U_l; U_k|_{k=1}^{l-1}, U_k|_{k=l+1}^{r}), l \in \{1, 2, \ldots, r\}$ is a function of U_l with given $U_k|_{k=1}^{l-1}$ and $U_k|_{k=l+1}^{r}$. According to the definition (13), the objective function is locally convex. The value of function (12) in the tth iteration approaches a local maximum as the projection matrices are updated. We can calculate $g_l(U_l^*)$ by maximizing $f_l(U_l)$ with the given $U_k|_{k=1}^{l-1}$ in the tth iteration and $U_k|_{k=l+1}^{r}$ in the $(t-1)$th iteration which are described in Steps 4–6 in Algorithm 1.

The alternating projection can be illustrated by a composition of r sub-algorithms defined as

$$\Omega_l : (U_l^*|_{l=1}^r) \mapsto \prod_{k=1}^{l-1} \times_k U_k \times U_l^* \times \prod_{k=l+1}^{r} \times_k U_k. \tag{14}$$

As discussed in [23] and [24], $\Omega \doteq \Omega_1 \circ \Omega_2 \cdots \circ \Omega_r$ is a closed algorithm and all sub-algorithms $g_l(U_l^*)$ increase the values of f, so it is clear that Ω is monotonic with respect to f. Therefore, the alternating projection method to optimize NTPCA converges. We can terminate the iteration procedure when the change of f between two successive iterations is sufficiently small.

III. Gabor Tensor Feature Extraction

The auditory system represents speech signals in both the temporal and spectral domain. The response of an auditory neuron in the primary auditory cortex (A1) can be described in terms of its spectro-temporal receptive field (STRF). In this paper, we employ 2-D Gabor functions to model the primary auditory cortical representation [29] in the spectro-temporal domain. The response of a population of cortical cells is represented in a high order feature space. The goal of our proposed method is to extract the intrinsic representation of auditory perception for building practical speech recognition systems.

A. Gabor-Based Multifactor Representation

As described in [30], the spectrum represents the responses of a population of cortical neurons. In the primary auditory cortex the auditory spectrum is decomposed into a more elaborate representation which contains the spectral and temporal modulation content. The neuron fires at its maximum rate for the input tones at its particular center frequency (CF). For a given time frame, the speech cortical representation is a higher order tensor [14], [30], [32], [33]. This tensor model has three factors: the center frequency f, the scales (spectral bandwidth) s, and the phase (local symmetry) ϕ. The scales describe the bandwidth of each response area along the tonotopic frequency axis. The phase describes the symmetry parameters of neuron response.

We employ 2-D Gabor functions to model the STRF of cortical cells in the auditory cortex [11] based on the observations [31], [33], [34] that response of those cells are tuned to localized spectro-temporal modulations. The 2-D Gabor function $g_{u,v}(f,t)$ is defined as

$$g_{u,v}(f,t) = g_{\bar{k}}(\bar{x}) = \frac{\bar{k}^2}{\sigma^2} \cdot e^{-(\bar{k}^2 \cdot \bar{x}^2 / 2\sigma^2)} \cdot \left[e^{i\bar{k} \cdot \bar{x}} - e^{-(\sigma^2/2)} \right] \tag{15}$$

where $\bar{x} = X(f,t)$ is the power spectrum X at frame t with component index of frequency f, $\bar{k} = k_v e^{i\phi}$ is a vector and controls the scale and direction of Gabor functions, where $k_v = 2^{-(v+2/2)} \cdot \pi$, $\phi = u(\pi/K)$. We manipulate the scale and direction of Gabor functions by changing parameters v and u, respectively, and K determines the total number of directions. The parameters u and v are linked and controlled by vector $\bar{k}$. We can obtain the temporal modulation and spectral modulation of speech by this Gabor function. Three typical examples of Gabor functions are shown in Fig. 1.

$$\begin{aligned} g_l(U_l^*) \doteq \arg\max_{U_l \geq 0, U_l \in S_l} f_l(U_l) = \arg\max_{U_l \geq 0, U_l \in S_l} & \frac{1}{2} tr \left(U_l^T \left(\sum_{i=1}^{n} \left[mat_l \left(\mathbf{X}_i \overline{\times}_l U_l^T \right) mat_l^T \left(\mathbf{X}_i \overline{\times}_l U_l^T \right) \right] \right) U_l \right) \\ & - \frac{\alpha_l}{4} ||I - U_l^T U_l||_F^2 - \beta_l \mathbf{1}^T U_l \mathbf{1} - \frac{\alpha_l}{4} \sum_{k=1, k \neq l}^{r} ||I - U_k^T U_k||_F^2 - \beta_l \sum_{k=1, k \neq l}^{r} \mathbf{1}^T U_k \mathbf{1} \end{aligned} \tag{13}$$

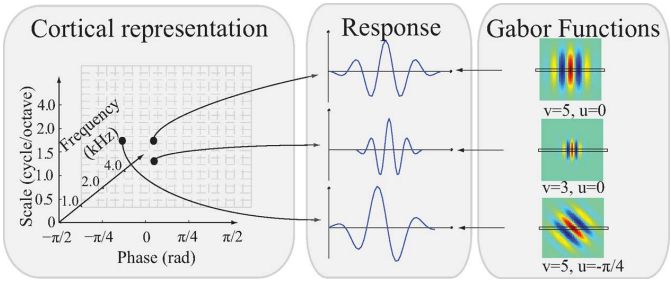

Fig. 1. Cortical representation of primary auditory cortex.

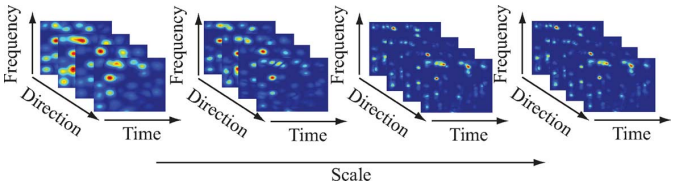

Fig. 2. Gabor tensor feature. The tensor structure has four independent factors: time, frequency, direction, and scale.

Fig. 1 represents the cortical representation with a primary auditory cortex neuron which has coordinates (x, s, ϕ). Three examples with different scale and direction parameters are given in Fig. 1. The first example with parameters $v = 5$, $u = 0$ describes the case that the response areas of neurons have a centered excitatory band with symmetrical inhibitory sidebands. The second example with parameters $v = 3$, $u = 0$ shows the band with of response areas shrinkages as parameter v decrease. The response areas in third example with parameters $v = 5$, $u = -\pi/4$ become asymmetric with stronger inhibitory sidebands above CF in one direction than below CF in the opposite direction. The response result can be simulated by 2-D Gabor functions as parameters u and v change.

Based on cortical representation in Fig. 1, we represent the power spectrum $X \in R^{N_f \times N_t}$ in given time window as a 4-order tensor $\mathbf{X} \in R^{N_f \times N_t \times N_v \times N_u}$ with four different factors: Frequency × Time × Scale × Direction. It is clearly seen that the cortical representation in Fig. 2 describes temporal-spectral patterns with different Gabor functions. Compared with the traditional spectral-temporal representation, the cortical representation is biologically inspired and gives more elaborate and abundant patterns for feature extraction. We calculate cortical representation by convolving the Gabor functions $g_{u,v}(f,t)$ with the power spectrum X and obtain a 4-order tensor $\mathbf{X} \in R^{N_f \times N_t \times N_v \times N_u}$ as shown in Fig. 2. For certain parameters u and v of Gabor functions, the Gabor filtering result $G_{u,v} \in R^{N_f \times N_t}$ is defined as

$$G_{u,v}(f,t) = |X(f,t) * g_{u,v}(f,t)| \tag{16}$$

where $X(f,t)$ is the power spectrum at frame t with component index of frequency f, $f = 1, \ldots, N_f$ and $t = 1, \ldots, N_t$, and $*$ is the convolution operator.

Thus, we transform the speech signal into a high-order representation of spectro-temporal modulations. The new representations $G_{u,v}$ with different parameters u, v describe multifactor characteristics of cortical representation. We use the Mel scale

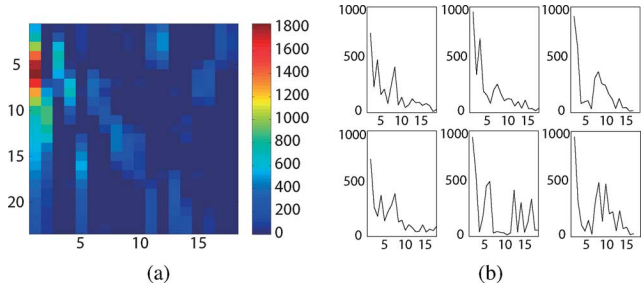

Fig. 3. Results of NTPCA applied to the clean speech data. (a) Projection matrix of NTPCA in spectro-temporal domain. (b) Samples of sparse coefficients (encoding) of feature vector (The x-axis is the dimension of the vector and the y-axis is the amplitude of the coefficients).

as the frequency warping scale to preserve the perceived frequency band. We define $G^{mel}_{u,v} \in R^{N_l \times N_t}$ as the feature set after Mel filtering:

$$G^{mel}_{u,v}(l,t) = \sum_{f=L_l}^{H_l} |v_l(f) G_{u,v}(f,t)| \tag{17}$$

where $v_l(f)|_{l=1}^{N_l}$ is a set of triangular filters which are linear approximately below 1 kHz and logarithm above, L_l is the lowest frequency, and H_l is the highest frequency. Finally we obtain the tensor-based representation $\mathbf{G}^{mel} \in R^{N_f \times N_t \times N_v \times N_u}$ as data preparation for learning the projection matrices of different subspaces.

B. Multifactor Analysis With Sparse Constraint

The auditory model transforms the speech signal into high-order multifactor feature space. We employ NTPCA to learn the projection matrices $U_l|_{l=1}^{4}$ in each factor (frequency, time, direction, and scale) from the Gabor-based cortical representation after the alternating projection optimization procedure. Intuitively, the projection matrices preserve the statistical characteristic of clean speech data and transform the cortical representation into a sparse feature space. The extracted tensor features characterize the elaborate spectro-temporal patterns of cortical representation and preserve principal components of multiple factors, which are superior to traditional subspace feature extraction methods.

In order to extract robust feature suitable for recognition, speech samples are first transformed to cortical representation $\mathbf{G}^{mel} \in R^{N_f \times N_t \times N_v \times N_u}$ and then are projected onto frequency, scale, and direction axes by matrices $U_f \in R^{N_f \times N_f^*}$, $U_v \in R^{N_v \times N_v^*}$, $U_u \in R^{N_u \times N_u^*}$ according to the cortical model in Fig. 1. We obtain the sparse tensor feature $\mathbf{S} \in R^{N_f^* \times N_t \times N_v^* \times N_u^*}$

$$\mathbf{S} = \mathbf{G}^{mel} \times_1 U_f^T \times_3 U_v^T \times_4 U_u^T. \tag{18}$$

An example of projection matrix in frequency domain is shown in Fig. 3(a) and most elements of this project matrix are near zero, which accords with the sparse constraints of NTPCA. Several samples of coefficients of feature vectors after projection illustrated in Fig. 3(b) also show the sparse characteristic of features. In Fig. 4, the statistics of the coefficients of the sparse tensor features (Clean, 0 dB, 10 dB), babble noise, white noise, and factory noise are presented for

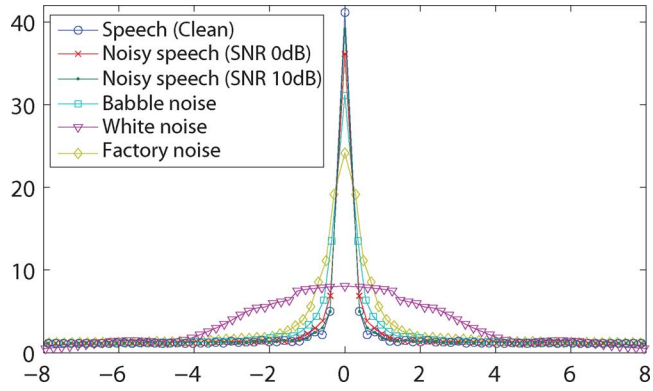

Fig. 4. Distribution of coefficients of the speech features (Clean, 0 dB, 10 dB), babble noise, white noise, factory noise. The x-axis is the coefficient s and the y-axis is the probability function $p(s)$.

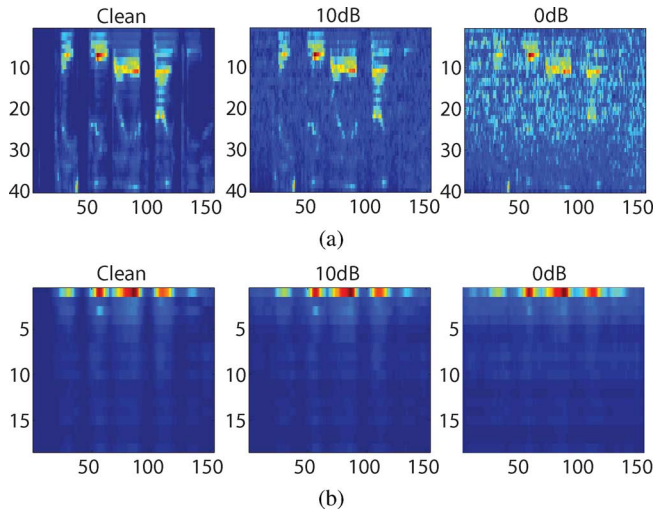

Fig. 5. (a) 40×153 standard mel spectrum and (b) 18×153 average sparse Gabor tensor feature under $u = 0$ for sentence "place blue by f nine please," before DCT under different SNR conditions (Clean, 0 dB, 10 dB).

comparison. We observe from computer simulations that the coefficient distribution of white noise is dense in the space of sparse speech representation, while the distribution of speech feature is sparse. Due to involving human speech in the babble noise, the feature distribution of babble noise is not as dense as white noise.

For comparison, Fig. 5 presents the standard Mel spectrum and the average sparse tensor features for sentence "place blue by f nine please" before DCT under different SNR conditions (Clean, 0 dB, 10 dB). In Fig. 5(a), Mel-based features with 40 filterbanks are shown and the degradation of the Mel spectrum in presence of additive factory noise is observed. The distortion of the Mel spectrum becomes severe with increase of noise intensity, resulting in performance degeneration in speech recognition by using MFCC features. From Fig. 5(b) we can find that the average sparse tensor features after projection maintain most spectral features from multiple subspaces compared with the features in clean conditions.

The sparse representation is efficient and coincided with the auditory neural coding [36]. Sparse coding theory [35] assumes that given a sound stimulus, only a few auditory neurons are active (nonzero elements) simultaneously. One can assume that

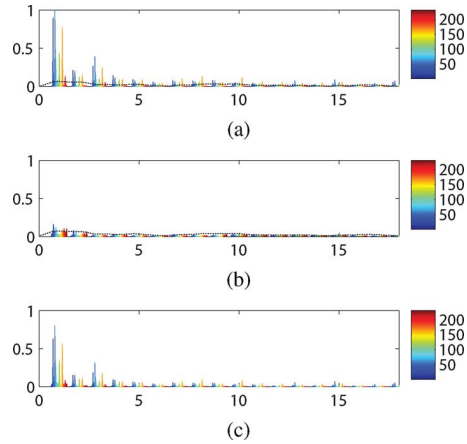

Fig. 6. (a) Feature coefficients distribution of clean speech. (b) Feature coefficients distribution of factory noise after transformation. (c) Feature coefficients distribution of speech mixed with factory noise (SNR $= 10$ dB). The x-axis is the dimension of feature vector. The y-axis is the amplitude of coefficients. The size of feature in (a)-(c) is 18×231. The x-axis include 18 bins and each bin include 231 coefficients of different samples.

the activity of neurons with small absolute values are purely noise and set them to zero, retaining just a few components with strong activities. The sparse constraint in the proposed method is closely related to the method of sparse coding shrinkage [35]. The constraint term $\mathbf{1}^T U_l \mathbf{1}, l = 1, \ldots, r$ will cause some elements of $U_l|_{l=1}^{r}$ to be exactly zero. As a result it can search for uncorrelated directions in which the components are as sparse (as in [35] super-Gaussian) as possible.

The sparse constraint results in feature robustness against noise because in sparse coding we are to find a projection subspace where the signal energy is only concentrated on a few components. When the speech signal is corrupted by additive noise, the Gaussian or weakly Gaussian noise usually has different distribution from speech, and therefore distribution of coefficients of noise in the space of sparse speech representation are almost uniformly spread over all components. Due to the sparse constraint, the coefficients of noise in the sparse speech representation will be suppressed when those coefficients are smaller than certain threshold. Therefore, the additive noise has little effect on the speech features. Thus, the features of clean speech are preserved in the sparse representation, while the noise components with dense distribution are suppressed.

Fig. 6 illustrates the sparse tensor feature distribution comparison of clean speech, factory noise and speech mixed with factory noise. The sparse tensor feature coefficients distribution of clean speech is presented in Fig. 6(a). Most elements in Fig. 6(a) is near to zero and the signal energy is concentrated on only a few components. Fig. 6(b) presents the feature coefficients distribution of factory noise in the feature space. Compared with sparse tensor features of clean speech, the coefficient distribution of factory noise is dense and uniform. The amplitude envelop of factory noise is given to compare the absolute value of noise signal with clean speech. From Fig. 6(c), we can find that the absolute value of feature coefficients is becomes smaller in the space of sparse speech representation compared with clean speech due to the shrinkage procedure, but the sparse tensor features of speech mixed with noise preserve most sparse components. This observation shows that the coefficients of noise

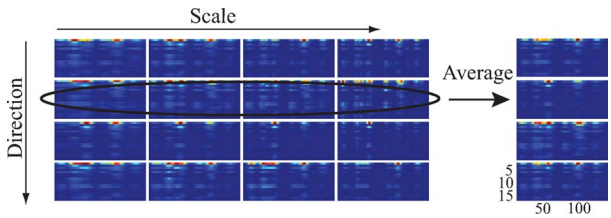

Fig. 7. Average scale features based on sparse tensor features.

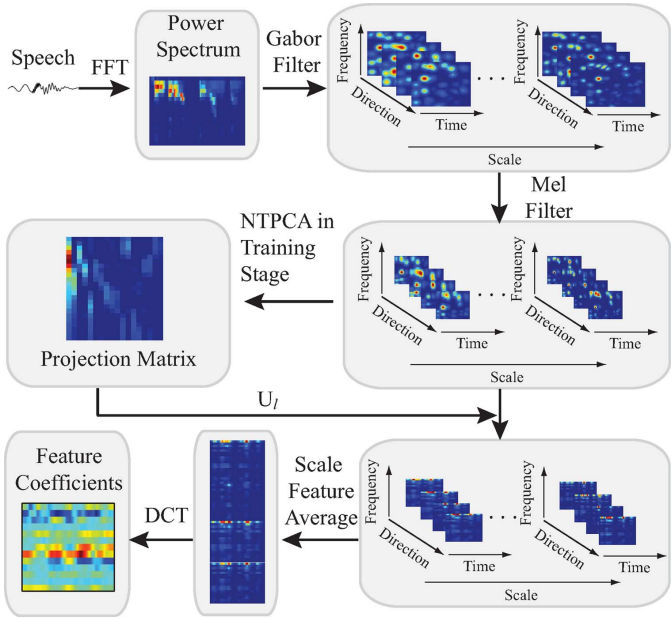

Fig. 8. Sparse Gabor tensor feature extraction framework.

with sufficiently small absolute values in the sparse representation are considered as disturbance of noise and will be discarded after projection.

Fig. 8 presents a speech feature extraction framework based on cortical representation and tensor analysis. We preform pre-emphasis on the speech signal and calculate the power spectrum by FFT. Then according to the cortical model, Gabor filtering and NTPCA are applied to transform the power spectrum into multiple feature subspaces. This Gabor-based sparse representation method is very similar to the image representation methods [37]. Here we employ the method proposed in [23] for the Gabor-based gait image representation to calculate the average scale features $G_{u,v}^{\text{avg}}(f,t)$, which are the average over scales of the Gabor-based sparse features $\mathbf{S}$ as shown in Fig. 7. Finally, we employ discrete cosine transform (DCT) on spectral feature vectors to reduce the dimensionality and de-correlate the feature components.

IV. Experiments Result and Discussion

In this section, we present computer simulation results of a speech recognition system using GTCC features in different noisy environments. Comparisons with MFCC feature and CMN, HLDA, Spectral Substraction, Temporal PCA, and NSPCA methods are also provided.

A. Grid Corpus

The performance of GTCC was tested on the Grid corpus which is created for research in speech separation and recognition. The total corpus consists of 17 000 sentences (500 from each of the 34 speakers). Sentences in the Grid corpus are six-word, fixed syntax utterances such as "bin blue at F 2 now." In this corpus, the vocabulary includes four verbs (bin, lay, place, set), four color choices (blue, green, red, white), four preps (at, by, in, with), 25 English letters ("W" is excluded due to its multi-syllabicity), ten digits (0 to 9), and four codas (again, now, please, soon). This recognition task is more difficult than digit- or letter-based corpora, for a more complex phone set.

B. Experimental Setup

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. At each window position, 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 four different scales and four different directions. Then the Gabor features were filtered by 40-channel Mel filterbanks to create the multifactor representation for tensor factorization.

We randomly selected 2000 sentences as a training dataset to learn projection matrices of data samples with tensor structure. The speech signals were transformed into tensor feature samples as the input for NTPCA. The GTCC feature vectors were obtained from the 26 sparse gabor tensor cepstral coefficients and delta ($\triangle$) and acceleration ($\triangle\triangle$) coefficients, which corresponded to a vector of 78 coefficients. We also provided the combination of 13 GTCC (without zeroth coefficient) and 13 MFCC and their $\triangle$ and $\triangle\triangle$ coefficients.

From the whole corpus, 8000 sentences were randomly chosen to train a speaker-independent recognizer using GTCC feature. The recognizer was a monophone-based system, where each word is a 18-state HMM with the probability density functions described by 3-Gaussian mixtures. 3600 sentences were mixed with six types of noise: white, babble, factory, leopard, m109, and destroyer operation room (each noise has 600 sentences), where noise samples were obtained from the Noisex-92 Dataset [38]. The SNR intensities were 15, 10, 5, and 0 dB for each noise. For comparison, the performance of MFCC, CMN, HLDA, spectral substraction, temporal PCA (TPCA) [8] with 39-order cepstral coefficients were tested. Similar to the feature extraction procedure of GTCC, NSPCA was performed on the frequency domain after Gabor filtering and the 26-order cepstral coefficients with $\triangle$ and $\triangle\triangle$ coefficients were provided for testing.

C. Evaluation Results

Fig. 9 shows the recognition accuracy of seven methods in different noise conditions. The speech recognition performance of GTCC was tested on six different noises with various SNR (15, 10, 5, and 0 dB). From Fig. 10, the above-mentioned methods achieve a high recognition performance for clean speech. Both GTCC and GTCC+MFCC provide acceptable accuracy in a clean condition.

As seen in Fig. 9, GTCC maintains appropriate recognition performance when noise reaches to signal levels (SNR reaching

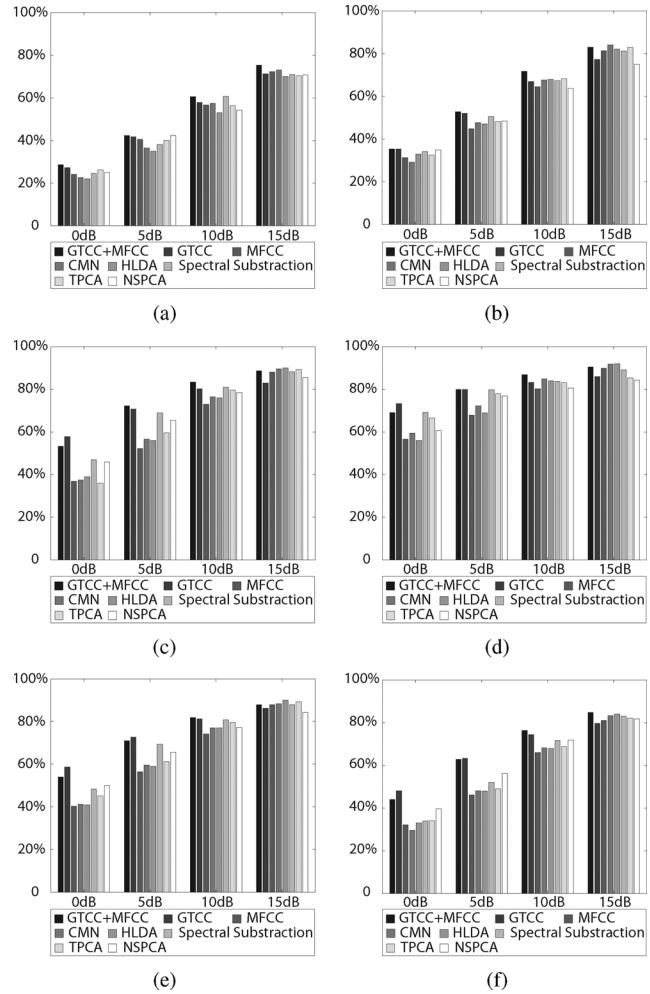

Fig. 9. Recognition accuracy in six noisy conditions (white, babble, factory, leopard, m109, and destroyer operation room) for Grid corpus. (a) White noise. (b) Babble noise. (c) Factory noise. (d) Leopard noise. (e) M109 noise. (f) Destroyer operation room noise.

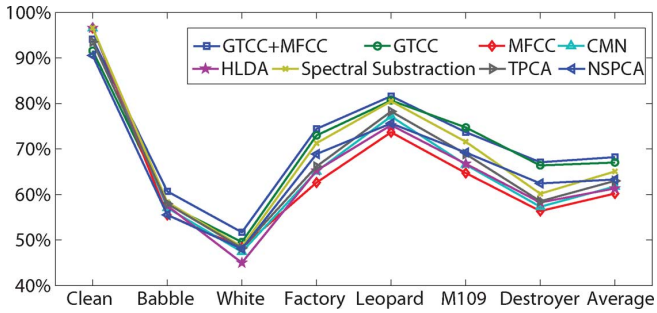

Fig. 10. Recognition accuracy in clean condition and six noisy conditions averaged over SNRs between 0–15 dB, and the overall average accuracy across all the noisy conditions, for GTCC, GTCC+MFCC and other features using Grid corpus mixed with additive noise.

0 dB). GTCC features perform significantly better in the presence of factory and M109 noise and slightly better in the presence of babble noise. The speech signal mixed with babble noise consists of other humans' speech signals, which corrupts the entire frequency bands and also shares the statistical properties of the reference signal. Then the performance of GTCC degrades although it is more robust than MFCC. For the other types of noise sources, their statistical characteristics are quite different from that of reference statistics, which GTCC features utilize to extract robust features. It is observed that GTCC features provide reasonable recognition performance especially for factory and M109 noise at very low SNR.

From the experimental result, the GTCC+MFCC give better performance than plain GTCC. This combination is similar to the overcomplete representation [39], which has greater flexibility in capturing structure in the data. Combing different basis functions into a single overcomplete basis would allow compact representations for widely types of signals. From Fig. 10 we can see that GTCC and GTCC+MFCC features show better average performance than the other features under different noisy conditions, indicating the potential of the new feature for dealing with a wider variety of noisy conditions.

D. Discussion

In this paper, we propose a multilinear framework for robust speech feature extraction. Compared with traditional subspace methods, NTPCA can preserve the intrinsic information in the natural structure of data through multifactor analysis. As described in Section II, multifactor analysis method may reduce the computational complexity since it processes the data sample in its natural structure with different factors alternatively.

Compared with other cortical models for feature extraction, our feature extraction model assumes that speech data in the feature space is sparsely distributed. The sparsity assumption is based on the fact that in sparse coding the energy of the signal is only concentrated on a few components. The learned projection matrices with sparse constraints preserve the statistical characteristics of clean training data. Therefore, the components consistent to statistical characteristics of clean speech data will be reserved during the feature extraction procedure. Since the noise generally has a different distribution from speech, the coefficients of noise in the feature space are usually distributed widely in most components. Then, the noise components with distributions different from speech will be suppressed through setting coefficients in the feature space to zero when those coefficients are smaller than a given threshold. The maximum variance criterion in each factor also ensures useful features in principal component subspace are maintained while the noise components are reduced.

Traditional spectral analysis methods such as MFCC have given an approximation of frequency integration of the auditory system in the 2-D spectro-temporal feature space. In this paper, we employ Gabor filters with different scales and directions to model the neuron response, which was motivated by cortical representation in the primary auditory cortex. As described in Section III, the feature space is a higher order tensor space with three independent modes. This model helps extract the robust features for recognition by investigating the intrinsic relations of different factors. Our proposed sparse Gabor feature is employed to describe the response characteristic of cortical neuron in temporal and spectral domain. These features contain more information on not only time dynamic characteristics but also the spectral features which are beneficial to robust feature extraction.

V. Conclusion

In this paper, we investigate the problem of speech feature extraction in noisy environments. Motivated by the auditory per-

ception mechanism, our method is focused mainly on the encoding of speech signals using a general higher order tensor structure. A new feature extraction method called NTPCA is developed for robust speech feature extraction from multiple feature subspaces based on tensor structure. The sparse constraints on NTPCA are employed to reduce the noise components and preserve useful information. The robust spectro-temporal features are extracted after multiple subspace projection. Experimental results demonstrate that the coding efficiency is improved compared with MFCC, CMN, HLDA, spectral substraction, TPCA, and NSPCA methods. It is believed that the proposed acoustic representations based on tensor structure can resemble auditory perception procedure to some extent.

APPENDIX

In this appendix, we give the detailed derivation of optimization problem (10). Based on the definition of tensor contraction, we have following equation:

$$[x \otimes x; (1)(1)] = \sum_{i=1}^{d} x_i x_i = tr(xx^T) \tag{19}$$

where $x \in R^d$ is a vector. Let $X \in R^{d\times n}$ denote a collection of column vectors $x_i|_{i=1}^n$ and $U \in R^{d\times k}$ a projection matrix. According to (19) and $\|A\|_F^2 = tr(AA^T)$ and $U^T A = A \times_1 U^T$, we obtain following equation:

$$\begin{aligned}\|U^T X\|^2 &= tr(U^T \left(\sum_{i=1}^{n} x_i x_i^T\right) U) \\ &= \sum_{i=1}^{n} tr(U^T x_i x_i^T U) \\ &= \sum_{i=1}^{n} \left[(x_i \times_1 U^T) \otimes (x_i \times_1 U^T); (1)(1)\right]. \end{aligned}\tag{20}$$

Based on equation $\|A\|_F^2 = tr(AA^T)$ and the definition of tensor contraction, we have the following equation for tensor $\mathbf{X} \in R^{N_1\times N_2\times\cdots\times N_r}$ and a set of matrices $U_k|_{k=1}^r \in R^{N_k\times N_k^*}$:

$$\begin{aligned}&\left[\left(\mathbf{X}\prod_{k=1}^{r}\times_k U_k^T\right)\otimes\left(\mathbf{X}\prod_{k=1}^{r}\times_k U_k^T\right); (1:r)(1:r)\right] \\ &= [(\mathbf{X}\overline{\times}_l U_l^T \times_l U_l^T) \otimes (\mathbf{X}\overline{\times}_l U_l^T \times_l U_l^T); (1:r)(1:r)] \\ &= \|\mathbf{X}\overline{\times}_l U_l^T \times_l U_l^T\|_F^2 \\ &= \|mat_l(\mathbf{X}\overline{\times}_l U_l^T \times_l U_l^T)\|_F^2 \\ &= \|U_l^T mat_l(\mathbf{X}\overline{\times}_l U_l^T)\|_F^2 \\ &= tr(U_l^T mat_l(\mathbf{X}\overline{\times}_l U_l^T) mat_l^T(\mathbf{X}\overline{\times}_l U_l^T) U_l). \end{aligned}\tag{21}$$

According to (20), we rewrite the optimization problem (8) as follows:

$$\begin{aligned}\bar{U} = \arg\max_{U\geq 0} \frac{1}{2}\sum_{i=1}^{n}\left[(x_i \times_1 U^T) \otimes (x_i \times_1 U^T); (1)(1)\right] \\ -\frac{\alpha_l}{4}\|I - U^T U\|_F^2 - \beta_l \mathbf{1}^T U \mathbf{1}. \end{aligned}\tag{22}$$

As the definition in Section II, $\mathbf{X}_i \in R^{N_1\times N_2\times\cdots\times N_r}$ is the ith training sample where $i = 1,\ldots,n$ and $U_l \in R^{N_l\times N_l^*}$ is the lth projection matrix where $l = 1,\ldots,r$. According to (22) and by replacing x_i with $\mathbf{X}_i$, we reformulate the nonnegative tensor principal component analysis (NTPCA) as the following optimization problem, as shown in (23) at the bottom of the page. The optimization problem defined in (23) does not have a close-form solution, so we employ the alternating projection method to find a solution. Based on (21), we decompose the optimization problem (23) into r different sub-problems, as shown in (24) at the bottom of the page. Then, we obtain the final optimization problem of NTPCA which is defined as (10) by introducing A_l, B_l and C_l to simplify (24).

$$\bar{U}_l|_{l=1}^r = \arg\max_{U_l\geq 0} \frac{1}{2}\sum_{i=1}^{n}\left[\left(\mathbf{X}_i\prod_{k=1}^{r}\times_k U_k^T\right)\otimes\left(\mathbf{X}_i\prod_{k=1}^{r}\times_k U_k^T\right); (1:r)(1:r)\right] - \frac{\alpha_l}{4}\sum_{k=1}^{r}\|I - U_k^T U_k\|_F^2 - \beta_l\sum_{k=1}^{r}\mathbf{1}^T U_k\mathbf{1}. \tag{23}$$

$$\begin{aligned}\bar{U}_l|_{l=1}^r &= \arg\max_{U_l|_{l=1}^r\geq 0} \frac{1}{2}\sum_{i=1}^{n} tr\left(U_l^T\left[mat_l\left(\mathbf{X}_i\overline{\times}_l U_l^T\right) mat_l^T\left(\mathbf{X}_i\overline{\times}_l U_l^T\right)\right] U_l\right) - \frac{\alpha_l}{4}\|I - U_l^T U_l\|_F^2 - \beta_l \mathbf{1}^T U_l \mathbf{1} \\ &\quad - \frac{\alpha_l}{4}\sum_{k=1,k\neq l}^{r}\|I - U_k^T U_k\|_F^2 - \beta_l\sum_{k=1,k\neq l}^{r}\mathbf{1}^T U_k\mathbf{1} \\ &= \arg\max_{U_l|_{l=1}^r\geq 0} \frac{1}{2} tr\left(U_l^T\left(\sum_{i=1}^{n}\left[mat_l\left(\mathbf{X}_i\overline{\times}_l U_l^T\right) mat_l^T\left(\mathbf{X}_i\overline{\times}_l U_l^T\right)\right]\right) U_l\right) - \frac{\alpha_l}{4}\|I - U_l^T U_l\|_F^2 - \beta_l \mathbf{1}^T U_l \mathbf{1} \\ &\quad - \frac{\alpha_l}{4}\sum_{k=1,k\neq l}^{r}\|I - U_k^T U_k\|_F^2 - \beta_l\sum_{k=1,k\neq l}^{r}\mathbf{1}^T U_k\mathbf{1} \end{aligned}\tag{24}$$

ACKNOWLEDGMENT

The authors would like to thank Associate Editor Prof. H. Li and anonymous reviewers for their constructive comments on this paper.

REFERENCES

[1] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[2] L. R. Rabiner and B. Juang, *Fundamentals on Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[3] A. Rosenberg, C.-H. Lee, and F. Soong, "Cepstral channel normalization techniques for HMM-based speaker verification," in *Proc. ICSLP*, 1994, vol. 4, pp. 1835–1838.

[4] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Jul. 1994.

[5] M. Berouti, R. Schwartz, J. Makhoul, B. Beranek, I. Newman, and M. A. Cambridge, "Enhancement of speech corrupted by acoustic noise," in *Proc. ICASSP'79*, 1979, vol. 4, pp. 208–211.

[6] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[7] K. Hermus, P. Wambacq, and H. Van hamme, "A review of signal subspace speech enhancement and its application to noise robust speech recognition," *EURASIP J. Appl. Signal Process.*, vol. 2007, no. 1, pp. 195–209, 2007.

[8] J. W. Hung and L. S. Lee, "Optimization of temporal filters for constructing robust features in speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 808–832, May 2006.

[9] J. W. Hung and W. Y. Tsai, "Constructing modulation frequency domain-based features for robust speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 3, pp. 563–577, Mar. 2008.

[10] X. Xiao, E. S. Chng, and H. Li, "Temporal structure normalization of speech feature for robust speech recognition," *IEEE Signal Process. Lett.*, vol. 14, no. 7, pp. 500–503, Jul. 2007.

[11] A. Qiu, C. E. Schreiner, and M. A. Escabi, "Gabor analysis of auditory midbrain receptive fields: Spectro-temporal and binaural composition," *J. Neurophysiol.*, vol. 90, no. 1, pp. 456–476, 2003.

[12] M. Kleinschmidt, "Localized spectro-temporal features for automatic speech recognition," in *Proc. 8th Eur. Conf. Speech Commun. Technol.*, 2003.

[13] T. Ezzat, J. Bouvrie, and T. Poggio, "Spectro-temporal analysis of speech using 2-D Gabor filters," in *Proc. Interspeech'07*, 2007.

[14] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 920–930, May 2006.

[15] J. Woojay and B.-H. Juang, "Speech analysis in a model of the central auditory system," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1802–1817, Aug. 2008.

[16] J. D. Carroll and J. J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart–Young" decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.

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

[18] R. Bro, "PARAFAC: Tutorial and applications," *Chemom. Intell. Lab. Syst.*, vol. 38, no. 2, pp. 149–171, 1997.

[19] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Anal. and A.*, vol. 21, no. 4, pp. 1253–1278, 2000.

[20] Y. D. Kim and S. Choi, "Nonnegative Tucker decomposition," in *Proc. CVPR'07*, 2007, pp. 1–8.

[21] A. Shashua and T. Hazan, "Non-negative tensor factorization with applications to statistics and computer vision," in *Proc. ICML'05*, 2005, pp. 792–799.

[22] A. Cichocki, R. Zdunek, S. Choi, R. Plemmons, and S. Amari, "Non-negative tensor factorization using alpha and beta divergences," in *Proc. ICASSP'07*, 2007, vol. 3, pp. 1393–1396.

[23] D. C. Tao, X. L. Li, X. D. Wu, and S. J. Maybank, "General tensor discriminant analysis and Gabor feature for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1700–1715, Oct. 2007.

[24] D. C. Tao, X. L. Li, X. D. Wu, and S. J. Maybank, "Tensor rank one discriminant analysis," *Neurocomputing*, vol. 71, pp. 1866–1882, 2008.

[25] L. D. Lathauwer, "Signal processing based on multilinear algebra," Ph.D. dissertation, Katholike Univ., Leuven, Belgium, 1997.

[26] C. A. Floudas and V. Visweswaran, "Quadratic optimization," in *Handbook of Global Optimization*. Dordrecht, The Netherlands: Kluwer, 1995, pp. 217–269.

[27] R. Zass and A. Shashua, "Nonnegative sparse PCA," in *Proc. Adv. Neural Information Process. Syst.*, 2007, vol. 19, pp. 1561–1568.

[28] Q. Wu and L. Q. Zhang, "Auditory sparse representation for robust speaker recognition based on tensor structure," *EURASIP J. Audio, Speech, Music Process.*, pp. 1–9, 2008.

[29] Q. Wu, L. Q. Zhang, and G. C. Shi, "Robust speech feature extraction based on Gabor filtering and tensor factorization," in *Proc. ICASSP'09*, 2009, pp. 4649–4652.

[30] K. Wang and S. A. Shamma, "Spectral shape analysis in the central auditory system," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 382–395, Sep. 1995.

[31] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Amer.*, vol. 118, pp. 887–906, 2005.

[32] R. Lyon and S. Shamma, "Auditory representation of timbre and pitch," *Auditory Comput.*, pp. 221–270, 1996.

[33] D. A. Depireux, J. Z. Simon, D. J. Klein, and S. A. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *J. Neurophysiol.*, vol. 85, pp. 1220–1234, 2001.

[34] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Commun.*, vol. 41, no. 2–3, pp. 331–348, 2003.

[35] A. Hyvärinen, "Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation," *Neural Comput.*, vol. 11, no. 7, pp. 1739–1768, 1999.

[36] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, no. 7079, pp. 978–982, 2006.

[37] C. Liu, "Gabor-based kernel PCA with fractional power polynomial models for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 5, pp. 572–581, 2004.

[38] The Noisex-92 Dataset, [Online]. Available: http://www.speech.cs.cmu.edu

[39] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Comput.*, vol. 12, no. 2, pp. 337–365, 2001.

Qiang Wu received the B.E. degree in computer science from Liaocheng University, Liaocheng, China, in 2002 and the M.E. degree in computer science from Shandong University, Jinan, China, in 2005. He is currently pursuing the Ph.D. degree in the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China.

His research interests include speech signal processing, brain–computer interfaces, machine learning, and neuroscience.

Liqing Zhang received the Ph.D. degree from Zhongshan University, Guangzhou, China, in 1988.

He was promoted to Full Professor in 1995 at South China University of Technology. He worked as a Research Scientist at the RIKEN Brain Science Institute, Japan from 1997 to 2002. He is now a Professor with Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests cover computational theory for cortical networks, brain signal processing and brain–computer interfaces, perception and cognition computing models, statistical learning, and inference. He has published more than 160 papers in international journals and conferences.

Guangchuan Shi received the B.E. degree in computer science and engineering from Shanghai Jiao Tong University, Shanghai, China, in 2008. He is currently pursuing the M.E degree at Shanghai Jiao Tong University.

His research interests include signal processing and perceptual computation.