ROBUST SPEECH FEATURE EXTRACTION BASED ON GABOR FILTERING AND TENSOR FACTORIZATION

Qiang Wu, Liqing Zhang, Guangchuan Shi

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

ABSTRACT

In this paper, we investigate the speech feature extraction problem in the noisy environment. A novel approach based on Gabor filtering and tensor factorization is proposed. From recent physiological and psychoacoustic experimental results, localized spectro-temporal features are essential for auditory perception. We employ 2D-Gabor functions with different scales and directions to analyze the localized patches of power spectrogram, by which speech signal can be encoded as a general higher order tensor. Then Nonnegative Tensor PCA with sparse constraint is used to learn the projection matrices from multiple interrelated feature subspaces and extract the robust features. Experimental results confirm that our proposed method can improve the speech recognition performance, especially in noisy environment, compared with traditional speech feature extraction methods.

Index Terms— tensor factorization, gabor filtering, feature extraction, speech recognition, auditory perception, acoustic noise

1. INTRODUCTION

The performance of speech recognition systems degrades in noisy conditions, which is a primary issue in utilizing such systems in real world. The degradation has been attributed to unavoidable mismatch between training and recognition conditions. Several methods have been proposed to reduce the effects of mismatch. Feature compensation techniques such as cepstral mean normalization (CMN), RASTA[1] have been developed for robust speech recognition. Feature extraction methods motivated by human auditory system[2] have been used to extract reliable features from speech signal, especially in noisy environments.

Recently the computational auditory nerve models attract much attention from both neuroscience and speech signal processing communities. Gabor STRF model[3] has been proposed to fit the auditory nucleus of interior colliculus by using spectral and temporal Gabor functions. Jeon[4] proposed a computational auditory central system model and interpreted various feature selection methods that parallel the computation of MFCC.

As a powerful data modeling tools, multilinear algebra of the higher order tensors has been proposed as a potent mathematical framework to manipulate the multiple factors underlying the observations. Currently common tensor decomposition methods include: (1) the CANDECOMP/PARAFAC model[5]; (2) the Tucker Model[6]; (3) Nonnegative Tensor Factorization (NTF) with nonnegative constraint on the CANDECOMP/PARAFAC model[7].

The work was supported by the National High-Tech Research Program of China (Grant No.2006AA01Z125), the National Natural Science Foundation of China (Grant No. 60775007) and the Science and Technology Commission of Shanghai Municipality (Grant No. 08511501700)

In this paper, 2D-Gabor functions are used to extract the spectro-temporal information, which employs multi-resolution wavelet over scales and directions to analyze the speech power spectrogram. Tensor analysis approach called NTPCA is derived by maximizing the covariance of data samples on tensor structure. The advantages of our method include following: 1) motivated by the human being auditory perception mechanism, multi-resolution spectro-temporal modulation with different scales and directions simulates the auditory cortical representation. The speech signal can be encoded as higher order tensor, which is beneficial to representing the perceptual information and improving robustness against noise. 2) sparse constraint on NTPCA enhances energy concentration of speech signal which will preserve the useful feature during the noise reduction. The Gabor tensor feature extracted by NTPCA can be further processed into a representation called Gabor Tensor Cepstral Coefficients(GTCC), which can be used as feature for speech recognition.

2. NONNEGATIVE TENSOR PRINCIPAL COMPONENT ANALYSIS

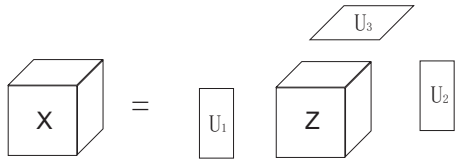


Fig. 1. Tucker model for tensor factorization

Multilinear algebra is the algebra of higher order tensors. A tensor is a higher order generalization of a matrix. Let $\mathbf{X} \in R^{N_1 \times N_2 \times \ldots \times N_M}$ denotes a tensor. The order of $\mathbf{X}$ is M. An element of $\mathbf{X}$ is denoted by $\mathbf{X}_{n_1,n_2,\ldots,n_M}$, where $1 \leq n_i \leq N_i$ and $1 \leq i \leq M$. The mode-i vectors of $\mathbf{X}$ are N_i-dimensional vectors obtained from $\mathbf{X}$ by varying index n_i and keeping other indices fixed. A tucker tensor factorization model is shown in Fig.1 where the core tensor $\mathbf{Z} \in R^{I_1 \times I_2 \times \ldots \times I_M}$ and $U_k \in R^{N_k \times I_k}, (k = 1, 2, \cdots, M)$. The Tucker model represents the data spanning the k-th modality by the vectors (loadings) given by the I_k columns of U_k such that the vectors of each modality interact with the vectors of all remaining modalities with strengths given by core tensor $\mathbf{Z}$. The details about multilinear algebra can be found in[6].

In [8], Zass proposed a nonnegative variant of the "Sparse PCA" called Nonnegative Sparse PCA. In this paper, we extend this model in the tensor structure. Let $\mathbf{X}_i$ denote the i-th training sample with zero mean which is a tensor, and U_k be the k-th projection matrix calculated by the alternating projection procedure. Here $\mathbf{X}_i (1 \leq i \leq n)$ are r-order tensors that lie in $R^{N_1 \times N_2 \times \cdots N_r}$ and

$U_k \in R^{N_k^* \times N_k}, (k = 1, 2, \cdots, r)$. We define nonnegative tensor principal component analysis by optimization problem(1) and use the alternating projection method, which is decomposed into r different optimization sub-problems:

$$\max_{U_l \geq 0 (l=1,\cdots,r)} \frac{1}{2} tr(U_l^T A_l U_l) - \frac{\alpha}{4} ||I - U_l^T U_l||_F^2 - \beta \mathbf{1}^T U_l \mathbf{1} + C_l, \tag{1}$$

where

$$A_l = \sum_{i=1}^{n} \left[mat_l \left(\mathbf{X}_i \overline{\times}_l U_l^T \right) mat_l^T \left(\mathbf{X}_i \overline{\times}_l U_l^T \right) \right], \tag{2}$$

$$C_l = -\frac{\alpha}{4} \sum_{k=1, k \neq l}^{r} ||I - U_k^T U_k||_F^2 - \beta \sum_{k=1, k \neq l}^{r} \mathbf{1}^T U_k \mathbf{1}, \tag{3}$$

In equation(1), $||A||_F^2$ is the squared Frobenius norm, $mat_d(\mathbf{X})$ is mode-d matricizing operator for tensor $\mathbf{X}$, $\overline{\times}_i = \prod_{k=1, k \neq i}^{r} \times_k$ is mode-d matrix product operator on each mode of tensor $\mathbf{X}$ except i-th mode, the second term relaxes the orthogonal constraint for common principal component disjoint, the third term is the sparse constraint, $\alpha > 0$ is a balancing parameter between reconstruction and orthogonality, $\beta \geq 0$ controls the amount of additional sparseness required. As described in[8], above optimization problem is a concave quadratic programming, which is an NP-hard problem. Therefore it is unrealistic to find the global solution of (1), and we have to settle with a local maximum. Here we give a function of u_{lpq} (the q row of the u_p column vector with index l) as the optimization objective,

$$f(u_{lpq}) = -\frac{\alpha}{4} u_{lpq}^4 + \frac{c_2}{2} u_{lpq}^2 + c_1 u_{lpq} + const, \tag{4}$$

where $const$ is the independent term of u_{lpq} and

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

$$c_2 = a_{lqq} + \alpha - \alpha \cdot \sum_{i=1, i \neq q}^{N_l} u_{lpi}^2 - \alpha \cdot \sum_{i=1, i \neq p}^{N_l^*} u_{liq}^2,$$

where a_{lij} is the element of A_l. Setting the derivative with respect to u_{lpq} to zero we obtain a cubic equation,

$$\frac{\partial f}{\partial u_{lpq}} = -\alpha u_{lpq}^3 + c_2 u_{lpq} + c_1 = 0, \tag{5}$$

We calculate the nonnegative roots of equation(5) and zero as the nonnegative global maximum of $f(u_{lpq})$. Table 1 lists the alternating projection optimization procedure for the Nonnegative Tensor PCA.

3. GABOR TENSOR FEATURE EXTRACTION

Inspired by recent physiological and psychoacoustic experimental results, much insight has been obtained from the measurements of so-called spectro-temporal response fields(STRF) of primary auditory cortex (AI) cells, which summarizes the way neuron cell responds to the stimulus. In this paper, we employ multi-resolution spectro-temporal modulation filters to model the primary auditory cortical representation. The power spectrum is encoded into a multilinear feature space by a population of cortical cells. The method we described is not biophysical in spirit, while rather it abstracts from the physiological and psychoacoustic experiments, which is likely to be relevant in the design of speech recognition system.

Table 1. Alternating Projection Optimization Procedure for NTPCA

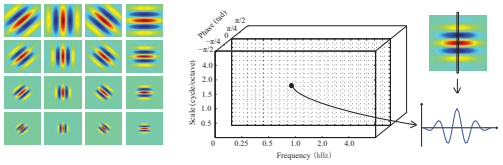
Input: Training tensor $\mathbf{X}_j \in R^{N_1 \times N_2 \times \cdots N_r}, (1 \leq j \leq n)$, dimensionality of the output tensors $\mathbf{Y}_j \in R^{N_1^* \times N_2^* \times \cdots N_r^*}$, α, β, maximum number of training iterations T, error threshold ε.
Output: projection matrix $U_l \geq 0 (l = 1, \cdots, r)$, $\mathbf{Y}_j$.
Initialization: Set $U_l^{(0)} \geq 0 (l = 1, \cdots, r)$ randomly, t=1;.
Step 1. Repeat until convergence {
Step 2. For l=1 to r {
Step 3. Calculate $\mathbf{A}_l^{(t-1)}$;
Step 4. Iterate over every entries of $U_l^{(t)}$ until convergence -Set the value of u_{lpq} to the global nonnegative maximizer of equation(4) by evaluating it over all nonnegative roots of equation(5) and zero;
}
Step 5. Check convergence: training stage of NTPCA if $t > T$ or update error $e < \varepsilon$
}
Step 6. $\mathbf{Y}_j = \mathbf{X}_j \prod_{l=1}^{r} \times_l U_l$

3.1. Gabor Functions

In [9] the neuronphysiological evidence indicates that the cells in the auditory cortex are tuned to localized spectro-temporal modulations. The STRF of these cortical cells [3] can be modeled by 2D Gabor functions. The 2D-complex Gabor function $g_{s,d}(t, f)$ is the product of a Gaussian envelope and a complex plane wave, defined as

$$g_{u,v}(t, f) = g_{\bar{k}}(\bar{x}) = \frac{\bar{k}^2}{\sigma^2} \cdot e^{-\frac{\bar{k}^2 \cdot \bar{x}^2}{2\sigma^2}} \cdot \left[e^{i\bar{k} \cdot \bar{x}} - e^{-\frac{\sigma^2}{2}} \right], \tag{6}$$

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



(a) Gabor functions (b) Cortical representation

Fig. 2. (a) The real part of Gabor functions for different scales and directions. (b) Cortical representation of primary auditory cortex.

3.2. Gabor-based Speech Tensor Representation

The primary auditory cortex analyzes the auditory spectrum into more elaborate representation and estimates the spectral and temporal modulation content. The spectrum is encoded by a population of cortical neurons, which are selective to different spectro-temporal modulation parameters. As the description in [10], the neural firing rates results are called cortical response. The output cortical

representation is higher order tensor structure, which are along three independent orders: the center frequency x, the scales(spectral bandwidth) s, the phrase (local symmetry) ϕ. The scales describe the bandwidth of each response area along the tonotopic frequency axis and phase denotes the symmetry.

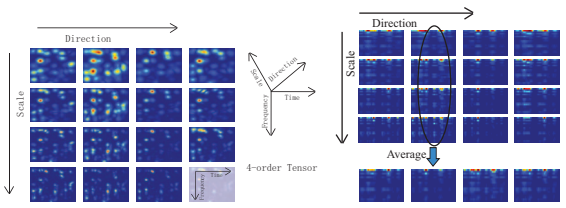
Above cortical representation based on tensor structure can be modeled by multi-resolution 2D Gabor transformation for the spectral pattern. From Fig.2(b) we can see the cortical representation with AI neurons that has its own (x, s, ϕ) coordinates. One example shows the cortical response in the case $\phi = 0$, which the neural response areas have a centered excitatory band that is symmetrically flanked by inhibitory side bands. While as ϕ increases above 0 rad, the response areas become more asymmetric with stronger inhibitory sidebands above CF in one direction and below the CF in the opposite direction. This corresponding result can be implemented by the 2D Gabor functions as direction parameter u changes.

The output of this cortical model is multiple dimensional array. In a given time window, the power spectrum $X(t, f) \in R^{N_t \times N_f}$ can be encoded as a 4-order tensor $\mathbf{X} \in R^{N_t \times N_f \times N_s \times N_d}$. From Fig.3(a) we can see clearly that the cortical representation has discriminative spectral patterns with different scale and direction parameters. This lie on the neuron response area.

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

$$G_{u,v}(t, f) = |X(t, f) \circledast g_{u,v}(t, f)|, \tag{7}$$

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



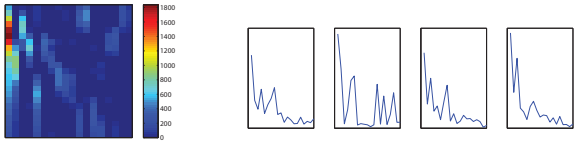
(a) Gabor tensor feature (b) average scale features

Fig. 3. (a) Gabor tensor feature. The rows show different scales and the columns show different directions for the power spectrum. (b) The average scale features based on sparse Gabor tensor feature.

3.3. Tensor Analysis and Sparseness Constraint

In order to extract the robust speech feature based on tensor structure, we transform the Gabor-based multi-resolution representation into multiple interrelated subspaces by NTPCA to learn the projection matrices $U_l, (l = 1, 2, 3, 4)$. 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. We employ the sparse localized projection matrix $U \in R^{d \times N_f}$ 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 via the following transformation:

$$S_{u,v}(t, f) = U * G^m_{u,v}(t, f), \tag{8}$$



(a) Projection Matrix (b) Samples for sparse feature vector

Fig. 4. (a)Projection Matrix. (b)Samples for sparse feature vector.

Fig.4(a) shows an example of projection matrix in spectro-temporal domain. From this result we can see that most elements of this project matrix are near to zero, which accords with the sparse constraint of NTPCA. Figure 4(b) gives several samples for coefficients of feature vector after projection which also prove the sparse characteristic of feature. The sparse constraint can make the feature robust based on the fact that in sparse coding the energy of signal is concentrated on a few components only, but the energy of additive noise remains uniformly spread on all the components. After sparse projection, the noise is reduced while the useful sparse information is not strongly affected.

Above Gabor-based sparse representation method is very similar to the image representation methods based on Gabor functions. However the computational cost of this method is high. Here we calculate the average scale features $G^{avg}_{u,v}(t, f)$ which are the sum over scales of Gabor-based sparse features. Fig.3(b) shows examples of the average scale features. At last we apply discrete cosine transform (DCT) on the average scale feature vectors to de-correlate the feature components. A diagram of feature extraction framework is shown in Fig.5.

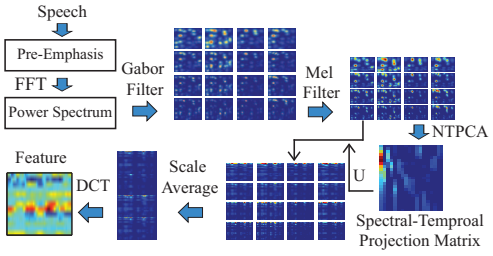


Fig. 5. The sparse Gabor tensor feature extraction framework.

4. EXPERIMENTS RESULT

In this section, we describe the evaluation results of speech recognition system using GTCC feature in the noisy environments. Com-

Table 2. Recognition accuracy in six noisy conditions averaged over SNRs between 0-15dB for GTCC and other features using Grid corpus mixed with additive noise.

(%)	GTCC	PLP	MFCC	CMN	HLDA
White	51.71	47.01	48.36	47.39	44.81
Babble	60.70	56.59	55.49	57.05	57.29
Factory	74.42	64.70	62.62	65.06	65.20
Leopard	81.62	74.97	73.68	77.15	75.17
M109	73.69	66.58	64.71	66.56	66.76
Destroyer	67.08	58.78	56.36	57.31	58.17

parisons with MFCC, PLP features and CMN, HLDA enhancement methods are also provided.

The performance of GTCC is tested on the Grid corpus. It was created for research in speech separation and recognition. The total corpus consists of 17000 sentences (500 from each of the 34 speakers). Sentences in the Grid corpus are 6-word, fixed syntax utterances such "bin blue at F 2 now". This recognition task is more difficult than the digit or letter based only corpora, for a more complex phone set.

The sampling rate of speech signal was 8kHz. 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 multiresolution Gabor-based feature 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 representation for tensor factorization.

We select 2000 sentences as training data randomly to learn projection matrices in each mode. The speech signals were transformed into tensor feature samples as the input for NTPCA. For the final feature set, the GTCC feature vectors for the evaluation were obtained from the 13 cepstral coefficients(without zero-th coefficient) combined with 13 mel cepstral coefficients and their delta ($\triangle$) and acceleration ($\triangle\triangle$) coefficients, which correspond to a vector of 78 coefficients.

From the whole corpus, 8000 sentences were randomly chosen to train a speaker-independent recognizer using GTCC feature. The recognizer was 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 noises white, babble, factory, leopard, m109 and destroyer operation room (600 sentences for each noise), where the noise samples were obtained from Noisex-92 Database. The SNR intensities were 15dB, 10dB, 5dB and 0dB for each noise. For comparison, the performance of PLP, MFCC, MFCC+CMN and MFCC+HLDA with 39-order cepstral coefficients are also tested.

For clean speech, the performance of both systems are comparable with high recognition rates where the word error rate is about 5% . Table 2 gives a average accuracy under different noisy conditions, respectively. GTCC features demonstrate significantly better performance in the presence of factory noise and slightly better performance in the presence of babble noise. For the babble noise, it 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 is reduced even though it is more robust than MFCC etc. While for other sources of noise, the characteristics of statistics are very different from that of reference statistics, which GTCC features can utilize to extract the robust features. The experimental results suggest that this auditory-based tensor representation feature is robust against the additive noise and suitable to the real application, indicating the potential of GTCC for dealing with a wider variety of noisy conditions.

5. CONCLUSION

In this paper, we considered the problem of speech feature extraction in noisy environment. Compared with traditional subspace learning methods, this study is mainly focused on encoding of speech signal into a general higher order tensor which simulated the cortical neuron responses model motivated by the auditory perception mechanism of human being. The sparse constraint on NTPCA helped to reduce the noise component and preserve the useful information. The discriminative and robust spectro-temporal feature was extracted after multiple subspace projection. Experiment result showed that the coding efficiency was improved compared with MFCC, PLP features and CMN, HLDA enhancement methods. This could be attributed to the ability of algorithm to find a better representation of the acoustic clues related to auditory perception model based on tensor structure.

6. REFERENCES

[1] L.R.Rabiner and B.Juang, *Fundamentals on speech recognition*. New Jersey: Prentice Hall, 1996.

[2] B.K.W.Mak, Y.C.Tam and P.Q.Li, "Discriminative auditory-based features for robust speech recognition", *IEEE Trans. Speech and Audio Process*, vol. 12, no. 1, pp. 27-36, 2004.

[3] A.Qiu, C.E.Schreiner and M.A.Escabi, "Gabor Analysis of Auditory Midbrain Receptive Fields: Spectro-Temporal and Binaural Composition", *Journal of Neurophysiology*, vol. 90, no. 1, pp. 456-476, 2003.

[4] J.Woojay and B.-H.Juang, "Speech Analysis in a Model of the Central Auditory System", *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 6, pp. 1802-1817, 2008.

[5] R.Bro, "PARAFAC: tutorial and applications", *Chemometrics and Intelligent Laboratory Systems*, vol. 38, no.2, pp.149-171, 1997.

[6] L.De Lathauwer, B.De Moor and J.Vandewalle, "A multilinear singular value decomposition", *SIAM Journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253-1278, 2000.

[7] A.Shashua and T.Hazan, "Non-negative tensor factorization with applications to statistics and computer vision", *Proceedings of the International Conference on Machine Learning, ICML'05*, pp. 792-799, 2005.

[8] R.Zass and A.Shashua, "Nonnegative Sparse PCA", *Advances in Neural Information Processing Systems 19*, pp. 1561-1568, 2007.

[9] T.Chi, P.Ru and S.A.Shamma, "Multiresolution spectrotemporal analysis of complex sounds", *The Journal of the Acoustical Society of America*, vol. 118, pp. 887-906, 2005.

[10] K.Wang, and SA.Shamma, "Spectral shape analysis in the central auditory system", *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 5, pp. 382-395, 1995.