International Conference of Electrical, Automation and Mechanical Engineering (EAME 2015)

A Primary Research on Gabor Tensor Sparse Features Representation for Whispered Speech Recognition

X.Q. Chen, H.M. Zhao, Y.B. Yu, H.W. Wu
School of Electronic and Information Engineering
Soochow University
Suzhou, P. R. China

Z. Liu
Department of Nanotechnology, Suzhou industrial Park
Institute of Services Outsourcing
Suzhou, P. R. China

Abstract **—Due to differences between normal and whispered speech, traditional feature performed poorly for whispered recognition. In this paper, a novel approach for whispered speech feature representation is proposed based on Gabor filtering and tensor factorization. The sparse feature is extracted by processing the data samples in tensor structure. The simulation results indicate that our proposed feature is able to improve the whispered speech recognition performance.**

Keywords-speech recognition; whispered speech; Gabor filtering; feature extraction

I. INTRODUCTION

Feature extraction is one of the important modules in speech recognition systems. The objective of feature extraction is to find discriminative features. 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]. These features are able to achieve acceptable accuracy for normal speech, while the recognition performance is degraded for whispered speech. Performance degradation is attributed to the difference between normal and whispered speech.

Whisper is a natural form of our daily communication. However, there are few researches on whispered speech. Reference[3] recognizes Japanese whispered speech by using Mel-frequency cepstral coefficient (MFCC) feature and Hidden Markov Models (HMM). It finally has a recognition rate of 68% which can be increased by 10% with maximum likelihood linear regression (MLLR) adaptive training approach. Reference[4,5] establishes a Chinese isolated word recognition system. The recognition rate of all these systems is lower than that of the normal speech for the special of whispered speech.

Recently, computational auditory neural models and sparse coding have attracted much attention in the societies of neuroscience and speech processing. The Gabor STRF model [6] was proposed to fit the auditory nucleus of inferior colliculus by using spectral and temporal Gabor functions. Reference [7] proposes a robust speech feature extraction method based on Gabor analysis and tensor factorization. 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. This method can keep the spectrum consistency between normal and whispered speech. In this paper, we used the Gabor tensor to express the structure characteristic of spectrum, experiments show that the proposed feature is more suitable for whispered speech recognition.

In the following part of this paper, a comparison for normal and whispered speech is discussed in Section II. the Gabor tensor feature extraction framework is described in Section III. Section IV gives experimental results and related analysis. Finally, Section V provides a discussion and conclusions.

II. NORMAL AND WHISPERED SPEECH

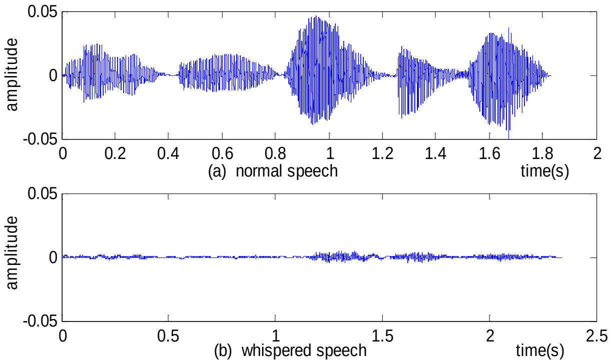

FIGURE I. WAVEFORMS EXAMPLE OF NORMAL SPEECH AND WHISPERED SPEECH SIGNALS.

Due to their different production mechanism, whispered speech is different from the normal speech. The main difference between whispered speech and normal speech is the complete absence of vocal fold vibration for whispered speech. In Fig.1 and Fig.2, the waveforms and spectrograms examples of the speech signal for both neutral and whispered speech are displayed. The intensity of whispered speech is significantly lower than that of neutral speech, the lack of voiced periodic components in whispered speech can also be observed. Due to these physiological differences in the production mechanism, the acoustic characteristics of whispered speech are different from those of normal speech. In spite of these differences, we found that the topological structure spectrum is similar between the two kinds of speech.

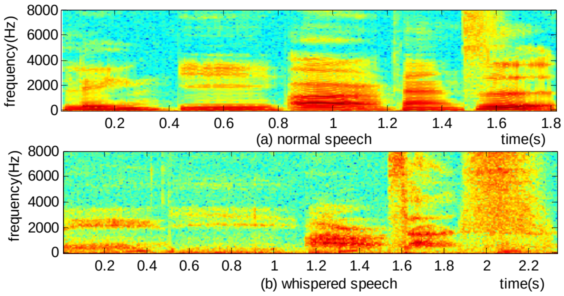

FIGURE II. SPECTROGRAMS EXAMPLE OF NORMAL SPEECH SIGNAL AND WHISPERED SPEECH SIGNAL.

III. FEATURE EXTRACTION

A. *Review of Gabor Tensor*

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 reference[7], 2-D Gabor functions is used to model the primary auditory cortical representation in the spectro-temporal domain. The response of a population of cortical cells is represented in a high order feature space. This method can extract the intrinsic representation of auditory perception of whispered speech.

Based on reference[7], the auditory spectrum is decomposed into a more elaborate representation which contains the spectral and temporal modulation content in the primary auditory cortex. For a given time frame, the speech cortical representation is a higher order tensor. This tensor model has three factors: the center frequency, the scales (spectral bandwidth), and the phase (local symmetry). The scales describe the bandwidth of each response area along the tonotopic frequency axis. The phase describes the symmetry parameters of neuron response. 2-D Gabor functions is used to model the STRF of cortical cells in the auditory cortex based on the observations 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:

$$g_{u,v}(f,t) = g_{\bar{k}}(\bar{x}) = \frac{\bar{k}^2}{\sigma^2}\cdot e^{-\left(\bar{k}^2\cdot\bar{x}^2/2\sigma^2\right)}\cdot\left[e^{i\bar{k}\cdot\bar{x}} - e^{-\left(\sigma^2/2\right)}\right]. \quad (1)$$

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 $\bar{k}$.

B. *Feature Extraction Framework*

We perform pre-emphasis on the speech signal and calculate the power spectrum by FFT firstly. Then according to the cortical model, Gabor filtering is applied to transform the power spectrum into multiple feature subspaces. Finally, we employ discrete cosine transform (DCT) on spectral feature vectors to reduce the dimensionality and de-correlate the feature components.

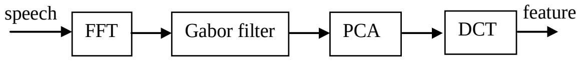

FIGURE III. SPEECH FEATURE EXTRACTION FRAMEWORK

IV. EXPERIMENTS OF WHISPERED RECOGNITION

A. *Recognition System*

This paper introduces a system based on Hidden Markov Model (HMM) for Chinese isolated word speech recognition. In the experiment, a Chinese speech database of isolated word composed of Arabia digital is established in laboratory environment. Normal and whispered speeches are both 40 groups recorded by 40 independent speakers with the format of 16 kHz/16 bit. In each group, each digital is spoken twice.

Each digital speech is modeled by 4-state left-to-right HMMs with 3 Gaussian mixtures per state. This experiment firstly extracts feature parameters of each number. The output probability of this sequence on each HMM is obtained. The number corresponding to the maximum output probability of HMM is the recognition result.

B. *Recognition Result*

From the experiment, different recognition rates are obtained by changing the numbers of states and mixtures of HMM, just as Table 1 and Table 2 shown.

TABLE I. RECOGNITION RATES OUTSIDE THE TRAINING SET (MFCC).

Train speech	Test speech	Recognition rate		
		4 states 3 mixtures	6 states 1 mixtures	6 states 3 mixtures
Normal	Normal	89.0%	90.0%	81.0%
	Whisper	**36.8%**	**23.6%**	**35.9%**
Whisper	Whisper	82.0%	81.0%	77.0%

TABLE II. RECOGNITION RATES OUTSIDE THE TRAINING SET (PROPOSED FEATURE).

Train speech	Test speech	Recognition rate		
		4 states 3 mixtures	6 states 1 mixtures	6 states 3 mixtures
Normal	Normal	83.0%	84.0%	78.0%
	Whisper	**62.6%**	**57.5%**	**60.6%**
Whisper	Whisper	81.0%	78.0%	79.0%

In table 1, where the training and test features are MFCC, the recognition of whispered speech is obviously lower than that of the normal speech. Especially when the training data is normal speech, the average recognition of whispered speech dropped to 32.1%. This is mainly caused by the mismatch between the acoustic characteristics in normal and whispered speech.

In table 2, where the training and test features are the proposed features, the recognition of whispered speech have no obvious reduced comparing the normal speech when training set and test set are agreement. When the training data is normal speech, the average recognition of whispered speech can keep the value of 60.23%. This is the significantly improvement, we think the proposed method can reserve the spectrum topological structure, which is the most similarities between the whispered speech and normal speech.

V. Conclusion

This paper compares the acoustic characteristics between Chinese normal and whispered speech. Aiming at the problem of the recognition rate decline of the whispered speech, we are committed to find a more suitable feature representation method for whispered speech. In this paper, a primary exploration is discussed using a general Gabor tensor structure to express the feature of whispered and normal speech. And experiments show that the proposed method can keep the common features of whispered speech and normal speech better. For further study to improve the recognition rate of whispered speech, more attention could be paid on the tensor parameters setting and sparse representation.

Acknowledgment

This work is supported by the national natural science foundation of China under grant No.61340004 and by the natural science foundation of Jiangsu province under grant No.BK20131196 and by the Key Laboratory of Advanced Information Science and Network Technology of Beijing (XDXX1305). It is also supported in part by the national natural science foundation of China under grant No.61372146 and No.61271360.

References

[1] S. B.Davis, P.Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *IEEE Transaction Acoustic, Speech, Signal Processing*, 28(4), pp. 357–366, 1980.

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

[3] T.Itoh, K.Takeda, & F.Itakura, Acoustic analysis and recognition of whispered speech. *International Conference on Acoustics, Speech and Signal Processing*, Orlando, Florida, USA, pp.389-392, 2002.

[4] L.L.Yang, W. Lin & B.L.Xu, Isolated word recognition in Chinese whispered speech. *Applied Acoustics*, 25(3), pp.187-192, 2006.

[5] W.Rong, Z.Tao, J.H.Gu & H.M.Zhao, Chinese whispered speech recognition based on improved LPCC and MFCC. *Computer Engineering and Application*, 43(30), pp.213-216, 2007.

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

[7] Q.Wu, L.Q.Zhang & G.C.Shi, Robust Multifactor Speech Feature Extraction Based on Gabor Analysis. *IEEE Transactions on Audio Speech and Language Processing*. 19(4), pp. 927-936, 2011.