DOMINANT SUBSPACE ANALYSIS FOR AUDITORY SPECTRUM

Lu Xugang, Li Gang, Wang Lipo

Nanyang Technological University , School of EEE, Workstation Resource Lab, Nanyang Avenue, 639798, Singapore

ABSTRACT

In hearing perception theory, spectral structure is a most important feature for speech perception, this spectral structure is not easy to be masked in noisy condition. So if this structure is extracted and enhanced, the representation will be much more robust. In this paper, we propose a new statistical dominant subspace analysis method for auditory spectrum based on SVD(Singular Values Decomposition) and signal subspace analysis method. The auditory spectrum can be decomposed into two subspaces, one is a dominant subspace, which is expanded by useful speech auditory spectrum , another subspace is sub-dominant subspace, which there is only noise information. So we analysis the auditory spectrum in the dominant subspace, the SNR will be increased. Thus this representation is much more robust.

1. COMPUTATIONAL AUDITORY MODEL AND AUDITORY SPECTRUM

Speech stimulation can be represented by auditory neural system in many stages. First, it can be decomposed into many frequency bands by basilar membrane, then after processed by inner hair cell and neural fibers, it's intensity is represented by neural firing rate. This neural impulse can be transformed to auditory central system, where it can be perceived by auditory cortex[1]. In this paper, all the processing parts are integrated using digital signal processing method, when speech signal is processed by this model, auditory feature can be gotten. The basic processing frame is as in Fig.1: the system is made up of six parts, that is, the high pass filtering of outer ear and middle ear, the band pass filtering of basilar membrane, nonlinear compression and half wave rectifying of inner hair cell, low pass filtering of neural fiber, energy detection of central system,

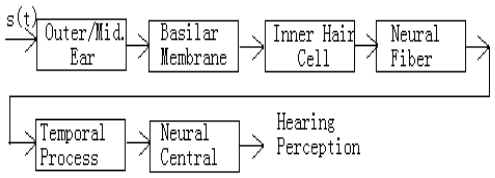

Figure1 Auditory model for speech signal processing

A mathematical model is designed to simulate this auditory function, as in Fig.2, a low pass filter is used to simulate the long temporal integration mechanism. The function of outer/middle ear can be simulated by high pass filter; band pass filters for basilar membrane; half-wave rectify for inner hair cell; low pass filter for neural fiber; energy detector and log compression for neural central, at last a DCT is used to get the feature vector.

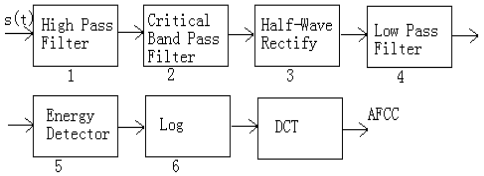

Figure2 The mathematical model for Figure 1

In these modules, short term adaptation and rapid adaptation of inner hair cell and neural fiber are not considered. Also, functions of temporal integration of neural central system and low pass filtering of neural fiber are integrated as a low pass filter. Energy detector is used for the intensity detection for each frequency channel. After processed by this model, a auditory feature is gotten. The feature can be used for training and testing. In this paper, we only focus on the auditory spectrum analysis, so the auditory spectrum can be got from the energy detector of Figure 2. Visual representations of FFT, LPC, and Auditory Spectrum are drawn for comparison in figure 3.(the spectrum of a Chinese sentence).

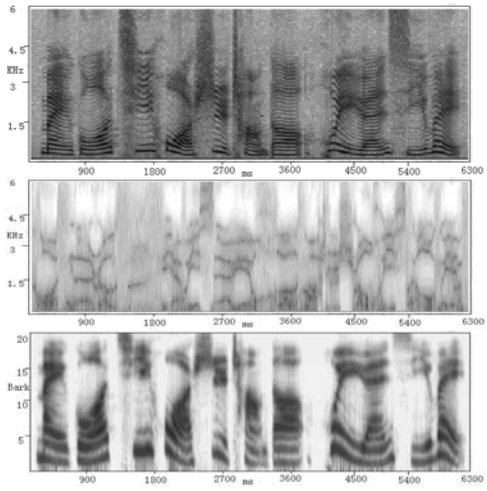

Figure3 Top is FFT spectrum, middle is LPC spectrum. Bottom is Auditory Spectrum(AS) .

From Fig.3, it is clear that AS(Auditory Spectrum) is wide band spectrum, FFT is narrow band spectrum. AS spectrum can be regarded as a smoothed spectrum of FFT

spectrum in hearing perception scale(in frequency domain). It is very clear that , speech representation by auditory system is a series of time-frequency patches, these patches are different from noise patch. We can regard the speech feature as a continuous time-frequency patch with regular structure. Noise patches is random and no-regular, so we hope subspace decomposition method can help use to separate noise and speech by this property.

2 SIGNAL SUBSPACE AND SVD

Signal subspace analysis method is widely used in digital signal processing and pattern recognition[2]. It is supposed that the useful information is only related with some lower dimensional subspaces, but noise is uniformly distributed in the whole measurement space (the whole Euclidean space). The subspace analysis method can decompose the whole measurement space into some useful subspaces, such as signal subspace and noise subspace or dominant subspace and subordinate subspace, then when the original feature is projected into the dominant subspaces, the dominant structure will be retained only, that is to say the subordinate feature(which including noise structure) will be reduced . From transformation view, we hope to find a new transform basis, the new feature gotten by this transformation can possess certain property, such as, each dimension of the feature vector is un-correlated or independent, etc. In this paper, we propose the subspace analysis method for the processing purpose.

Singular Value Decomposition (SVD) is a very useful method for matrix structure decomposition. we give some useful formulas here which can be used later. Suppose the data matrix is $X \in R^{m\times n}$, there exist orthogonal matrices:

$U = (u_1,...,u_m) \in R^{m\times m}$

$V = (v_1,...,v_m) \in R^{n\times n}$

satisfying:

$$X = U\Sigma V^T = \sum_{i=1}^{q} u_i \lambda_i v_i^T \qquad (1)$$

where $\Sigma = diag(\lambda_1,....,\lambda_q) \in R^{m\times n}$,

$\lambda_1 > \lambda_2 > ... > \lambda_q$, $q = \min(m,n)$

From this formula, it is clear that u_i can be regarded as the basis for the column vectors of X , the column vectors of X can be represented as:

$$X_{\bullet j} = \sum_{i=1}^{q} \lambda_i v_{ij} u_i \qquad (2)$$

This form can be explained as: the column vector of the data matrix can be represented by the combination of the weighted basis vector u_i , where the weighted coefficient is $\lambda_i v_{ij}$. While v_i can be regarded as the basis for the row vectors of X

$$X_{j\bullet} = \sum_{i=1}^{q} \lambda_i u_{ij} v_i^T \qquad (3)$$

This form can be explained as: the row vector of the data matrix can be represented by the combination of the weighted basis vector v_i , where the weighted coefficient is $\lambda_i u_{ij}$. In fact, all these representations can be regarded as finding an orthogonal transform A, that diagonalizes R_y :

$y = Ax, y \in R^n, x \in R^m, A \in R^{n\times m}$,

Generally , $n > m$, this is also a method of dimension reduction. For rank deficient data matrix, the SVD can be represented as partition SVD as:

$$X = U\Sigma V^T = (U_1 \quad U_2)\begin{pmatrix}\Sigma_1 & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}V_1^T\\ V_2^T\end{pmatrix} \qquad (4)$$

where $U_1 = (u_1,...,u_p) \in R^{m\times p}$

$,V_1 = (v_1,...,v_p) \in R^{n\times p}$,

$\Sigma_1 = diag(\lambda_1,....,\lambda_p) \in R^{p\times p}$

In noise condition, the data matrix is contaminated by noise, the rank of the data matrix maybe be full rank, because noise distribute in the whole Euclidean space, then the SVD can be represented as:

$$X = U\Sigma V^T = (U_1 \quad U_2)\begin{pmatrix}\Sigma_1 & 0\\ 0 & \Sigma_2\end{pmatrix}\begin{pmatrix}V_1^T\\ V_2^T\end{pmatrix} \qquad (5)$$

where $U_1 = (u_1,...,u_p) \in R^{m\times p}$

$,V_1 = (v_1,...,v_p) \in R^{n\times p}$,

$\Sigma_1 = diag(\lambda_1,....,\lambda_p) \in R^{p\times p}$

$\Sigma_2 = diag(\lambda_{p+1},....,\lambda_n) \in R^{(m-p)\times(n-p)}$

the singular values in Σ_1 are much larger than those in Σ_2 , we can explain that the larger singular values in Σ_1 are the signal structure, the much smaller singular values in Σ_2 are the noise structure(Strictly speaking, the noise fills the whole matrix structure). So the signal structure can be reconstructed by a low rank matrix approximation as:

$$X_S = U_1 \Sigma_1 V_1^T \quad (6)$$

Apparently, in the approximation matrix (6), the main structure is kept.

For a pattern space, we call the main structure space of the pattern as dominant subspace, it is expanded by the eigenvectors with larger singular values of the observation matrix gotten by SVD. New feature or new pattern matrix can be gotten by projecting the observation feature into this subspace. According to the property of the SVD, each dimension of the new feature will be un-correlated. Also it is easy to see, the left singular eigenvectors are used for row decorrelation(for spatial decorrelation), and the right singular eigenvectors are used for column decorrelation(for temporal decorrelation)

Now, let's taste some formulas of projection matrix and project coefficients. From the analysis above, the U_1 expands the column signal subspace for data matrix, and V_1 spans the row signal subspace of the data matrix. Then the column orthogonal subspace projection matrix is:

$$P_{U_1} = U_1 U_1^T = \sum_{i=1}^{p} u_i u_i^T \quad (7)$$

And the row orthogonal subspace projection matrix is:

$$P_{V_1} = V_1 V_1^T = \sum_{i=1}^{p} v_i v_i^T \quad (8)$$

So for an observed data vector z, it can be represented by the projection of the data into the subspace as:

$$\tilde{z} = \sum_{i=1}^{p} a_i u_i = \sum_{i=1}^{p} <z, u_i> u_i = P_{U_1} z \quad (9)$$

In transformation view, transformation coefficients are formulate as:

$$a_i = <u_i, z> = u_i^T z \quad , i = 1,2,...,p$$

or

$$a_i = <v_i, z> = v_i^T z \quad , i = 1,2,...,p \quad (10)$$

The signal subspace analysis method can be used both on the original speech data and on transformed data. For the original speech data, the data matrix is always Toeplitz form, such as for a speech frame $x = (x(1), x(2), ..., x(K))^T \in R^K$, the data matrix is:

$X = Toeplitz(x) \in R^{m \times n}$ where $K = m + n + 1$ .But in this paper we focus on the transformed matrix structure, the data matrix is formed by auditory spectrum.

3 .SUBSPACE ANALYSIS FOR AUDITORY SPECTRUM

From the first part of this paper, speech stimulation can always be represented by a temporal-spatial pattern according to the function of the auditory system, we can also regarded the temporal-spatial pattern as a time-frequency pattern with the corresponding resolution of auditory system(such as the temporal integration of time window, and bark frequency resolution). In hearing perception, the dominant parts of perception are larger time-frequency patches which can be integrated together by auditory system to give a perception. Such as formants transition structure, pitch transition structure, etc. We chose the primary auditory spectrum as the data matrix, the row is time dimension while column is frequency dimension. The data matrix can be decomposed into dominant part and subdominant part as formula (5), the reconstruction by the dominant subspace for the data matrix is as formula(6), while the minor time-frequency patches will be discarded during the reconstruction. The following figure4 is an example of a seven-syllable Chinese word, top is the original auditory time-frequency pattern contaminated by noise, the data matrix is $X \in R^{240 \times 172}$, that is 240 frequency channels and 172 time frames, bottom is the subspace approximation spectrum with only 10 dimension retained.

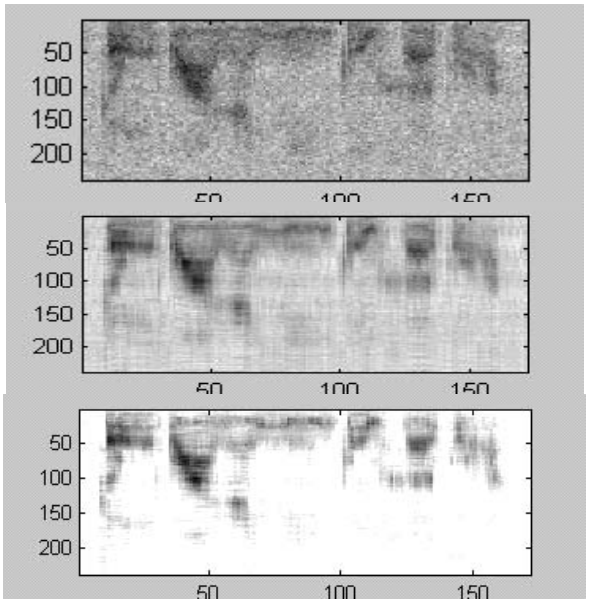

Figure4 Top is the noisy auditory spectrum, middle is the subspace approximation auditory spectrum by using only 10 largest singular values and corresponding eigenvectors, bottom is subspace approximation by discarding those minor time-frequency patches. X-axis is the time frame, Y-axis is the frequency channel

From figure4, it is very clear that the subspace approximation is a smoothed approximation of the

original matrix, if use a shrinkage function to get rid of those minor time-frequency patches, the subspace time-frequency structure is those larger time-frequency patches will be kept only. Figure5 is the normalized log singular spectrum of the data matrix(auditory spectrum).

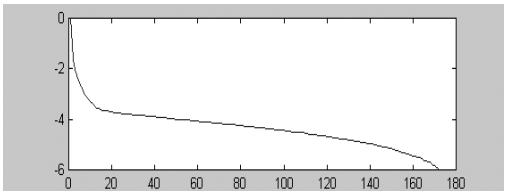

Figure5 The normalized log singular values spectrum of the data matrix structure, X-axis is the number of the singular values

It is very clear that the singular values decreases sharply, that is to say, the dominant structure of the data matrix can be kept well only by a few dimensions (with those larger singular values). So the dominant subspace of this data matrix can be expanded by the vectors corresponding to the larger singular values. The signal projected onto the subspace can keep the main structure. Figure 6 is the ratio of the signal energy and noise energy in different dimension subspaces.

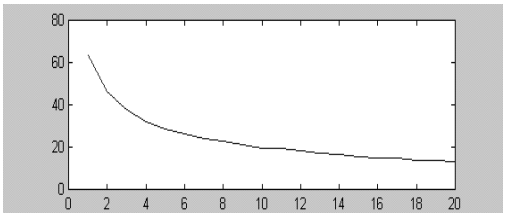

Figure 6 The ratio of Signal Energy/ Noise Energy , X-axis is the subspace dimension number(corresponding with those largest singular values), Y-axis is the Energy ratio

It is clear from figure 6, that with the increasing of the subspace dimension, the SNR is decreasing. Some experiments have been done based on these subspace analysis, but it shows that , the robustness is not improved so much after the subspace analysis. In fact, when speech is perceived by hearing system, it is perceived by some certain time and frequency resolution, the spectrum we used by computational auditory model is enough in time and frequency scales. That is to say, further low filtering of the auditory spectrum in time and frequency domain will not give any help for recognition. But some certain filtering in time or frequency domain can help to reduce some kinds of noise, such as acoustic channels or slow varying distortion. But shrinkage of the subspace spectrum can improve the robust, because the shrinkage can reduce the minor time-frequency patches. The following curve is the recognition rate .The SNR is defined as:

$$SNR = 10\log_{10}(x^2 / n^2) \qquad (11)$$

where x is the speech signal, n is the additive white noise. It is clear that for auditory spectrum, the robustness normalized AFCC with full dimension is better.

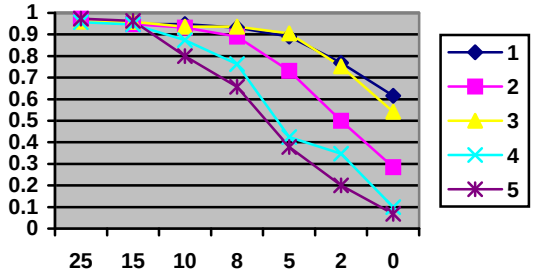

Figure 7 recognition rate in different condition, x-axis is the SNR, y-axis is the recognition rate, 1 is normalized AFCC with shrinkage, 2 is normalized AFCC with subspace dimension 11 and shrinkage, 3 is normalized AFCC without shrinkage, 4 is the AFCC without any normalization and shrinkage, 5 is MFCC

4. CONCLUSION

From the analysis above, the matrix structure given by the auditory spectrum can represent the time-frequency patch , also our hearing system can always trace these time-frequency patches to percept the stimulation. Subspace based approximation matrix keeps all the prominent time-frequency patches. In noisy condition, the spectral structure will be distorted, there will be some small ripples given by noise, time-frequency patches given by noise is smaller and not regular. That is to say, noise can distort the minor time-frequency patches of speech, but the larger time-frequency structure will not be distorted, such as the formants transition structure. In the subspace matrix approximation, the main time-frequency structure is kept well only with a basis vectors corresponding the larger singular values. Those minor time-frequency structure is discarded because of the discarding of the basis corresponding to those smaller singular values. In traditional speech processing method, this concept is already in use, such as Mel or Bark scale for frequency, and time frame integration for time dimension, that is the auditory resolution subspace measurement.

REFERENCE

1. Seneff S., A joint synchrony/mean-rate ,odel of auditory speech processing. Journal of Phonetics, Vol.16, 1988,pp.55-76
2. P.S.K., hansen, Signal subspace method for speech enhancement, 1997