TIME-FREQUENCY REPRESENTATION OF CHINESE SPEECH SIGNAL

Bowen ZHOU, Limin DU
Laboratory for Interactive Information Systems
Institute of Acoustics, Chinese Academy of Sciences
Tel: (8610)62627570 Fax: (8610)62629250 E-mail: zbw@farad.ioa.ac.cn

ABSTRACT

This paper first describes the theory and the implementation of Wigner Distribution (WD), a kind of joint time-frequency representation. We apply Wigner Distribution to the representation of Chinese speech signal. Ridges detection algorithm for extracting formant models from the smoothed joint time-frequency representations is also described and some experimental results are provided. From the experimental results, we can see the advantages of joint time-frequency representations over the traditional short-time Fourier technology.

1. INTRODUCTION

Time-frequency representations are multidimensional transformations that indicate the joint time frequency content of a signal. Representations such as the wavelet, the short-time Fourier transform (STFT) and the Wigner Distribution (WD) have been proven to be powerful tools for signal analysis and processing.

In the speech processing area, formant models are key information for both speech recognition and speech synthesis. The formant locations specify the general vowel quality, r-coloring and roundness, while the formant transitions between consonants and vowels play an important role in consonant identification ([2]). STFT was widely used to extract the formant models, based upon the assumption of short-time stationary property of speech signal. In fact, speech signal, especially the component of the speech that related to the response of vocal is obviously non-stationary. For instance, the second formant of some syllables in Chinese speech can vary with time at the speed of dozens of Hertz per millisecond. In this case, STFT is not an efficient tool to depict the formant models of these syllables.

Considering that speech signal is a typical kind of non-stationary signal， we apply the joint time-frequency representation, Wigner Distribution, to the representation of the Chinese speech signal. Ridges detection algorithm for extracting formants from joint time-frequency representations is also advanced and some experimental results are provided. From the experimental results, we can see the advantages of joint time-frequency representations over the traditional Short-time Fourier technology.

2. CHINESE SPEECH ANALYSIS WITH WIGNER DISTRIBUTION

Various ways have been used to express signal energy as a joint function of time and frequency. As an example, Wigner distribution (WD) is currently quite popular in the signal processing literature.

2.1 Introduction to Wigner distribution, pseudo-Wigner distribution

Given signal x (t), the WD of x (t) is defined as:

$$WD_x(t,\omega)=\int_{-\infty}^{\infty} x(t+\tau/2)x^*(t-\tau/2)e^{-j\omega\tau}d\tau \quad (1)$$

Here, '*' denotes the conjugate operator. WD shows many superior mathematical properties such as real-value, shift invariance, preserving the marginal distributions and the finite support properties. The

Wigner distribution, however, does not satisfy the properties of positivity or superposition (it means that the time-frequency representation of a multi-component signal should be a simple superposition of its components.). In other words, the Wigner distribution suffers from cross terms to which there can not be attributed much physical significance. Cross terms are undesirable and should be checked out for real application.

For infinite signal, for example, speech, the Wigner distribution is changed to following form, named as pseudo Wigner distribution (PWD),

$$PWD_x(t,\omega)=\int_{-\infty}^{\infty}x(t+\tau/2)x^*(t-\tau/2)\left|w(\tau/2)\right|^2e^{-j\omega\tau}d\tau \quad (2)$$

Where, $\omega(t)$ denotes an even symmetrical sliding window that intercepts the infinite signal.

2.2 The implementation of PWD

Toward the real application of Wigner distribution, we face the problems of how to calculate the discrete PWD (DPWD) and how to alleviate the effects of cross-terms introduced by the Bi-linear nature of the Wigner distribution.

DPWD is periodical in digital frequency domain with the period of π other than 2π. As a result, it enlarges the amount of computation for DPWD ([1]). One common method is to calculate DPWD by FFT. Because of the real-value property of WD, however, FFT, as a kind of complex value transform, is not the most efficient way. On the other hand, Hartley transform is a real-value transform ([3]), which is defined as:

$$H(k)=\frac{1}{\sqrt{N}}\sum_{n=0}^{N-1}f(n)cas(kn2\pi/N) \qquad (3)$$

Where,

$$cas(x)=\cos(x)+\sin(x)$$
$$k=0,1,...,N-1$$

More important, Hartley transform has mature fast algorithm similar to FFT. Therefore, computing DPWD by FHT is more efficient than FFT in this case ([1]).

Another important issue for real application of WD is to restrain the disturbance of cross terms, as stated before. There are many methods have been proposed for this task and we adopt the method of two dimensional convolutional smoothing in time-frequency domain to mitigate the effects of cross-terms:

$$\begin{aligned}WD_x'(t,\omega)&=\frac{1}{4\pi^2}\iint WD_x(t-\mu,\omega-\xi)\varphi(\mu,\xi)\,d\mu d\xi\\&=WD_x(t,\omega)**\varphi(t,\omega)\end{aligned} \qquad (4)$$

Where, $\varphi(t,\omega)$ is called as smoothing kernel function.

2.3 Applying PWD to Chinese speech signal analysis

For speech signal analysis, we select a two-dimensional Gaussian function to smooth the PWD and suppress the effects of cross-terms. The kernel function $\varphi(t,\omega)$ is:

$$\varphi(t,\omega)\propto e^{-t^2/4\sigma_t^2}e^{-\omega^2/4\sigma_\omega^2} \qquad (5)$$

For a specified group of σ_t,σ_ω, time-frequency scale is determined by it. So it is very important to choose the values of σ_t,σ_ω, which decide the smoothing effects. The Wigner distribution of speech signal is corrupted not only by the disturbance of cross-terms, but also by the interference of voiced excitation (see figure1 (b)). Choosing the value of σ_t corresponding to pitch period，and the value of σ_ω matching with fundamental frequency can suppress the disturbance effectively (see figure1 (c)).

3. FORMANT EXTRACTION FROM TIME-FREQUENCY REPRESENTATION

In this part, we will focus on how to find time-

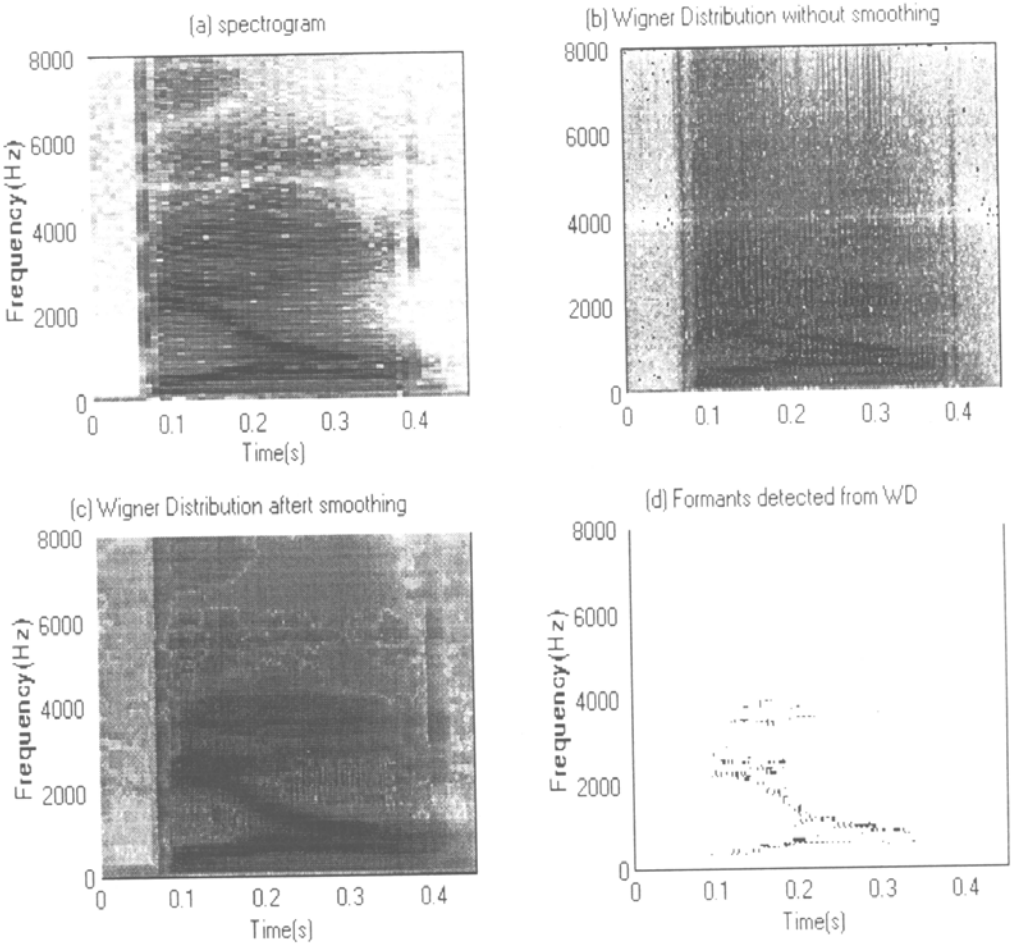

Figure 1. The result of syllable “biao1”, which is generated by an adult male
(a.) The spectrogram from STFT;
(b.) The time-frequency representation from WD, no smoothing;
(c.) The smoothed WD by a two-dimension Gaussian function;
(d.) Formant loci extracted from (c.), using ridge detection algorithm

frequency ridge due primarily to the formants. By that, we can produce a symbolic representation, named as schematic spectrogram ([2]), which captures the acoustically relevant features present in the joint time-frequency energy representation. It will be useful for finding onsets, offsets and bursts from time discontinuities, useful for finding the formants and perhaps channel resonance, also useful for formant and channel analysis.

After getting the smoothed pseudo Wigner distribution $F(t, \omega)$, we apply an algorithm, named as Ridge Detection (RD) to extract the schematic spectrogram from the smoothed time-frequency surface ([2]). The algorithm is described as follow:

a. For every point of $F(t, \omega)$, computing $\nabla F = (\frac{\partial F}{\partial t}, \frac{\partial F}{\partial \omega})$

b. Computing the Hessian matrix of $F(t, \omega)$, as

$\begin{pmatrix} \frac{\partial^2 F}{\partial t^2} & \frac{\partial^2 F}{\partial t \partial \omega} \\ \frac{\partial^2 F}{\partial \omega \partial t} & \frac{\partial^2 F}{\partial \omega^2} \end{pmatrix}$. Assigning ξ as the eigenvector that corresponds to k, the minimum eigenvalue. Then computing $gdcF = \frac{\xi}{|\xi|}$

c. For every point of time-frequency representation, if it satisfies

$$\nabla F \bullet gdcF = 0 \text{ And K<0,} \qquad (6)$$

then we conclude that the point is part of the formant ridge.

4. EXPERIMENTAL RESULT ANALYSIS

We have applied this method to many typical Chinese syllables. Figure 1 gives the result of “biao1”, a syllable articulated by an adult male, as an example of these results. This utterance has rapid F2 motion, which makes it useful as an example of non-stationary behavior in speech. Part (a.) shows the spectrogram calculated from STFT. Contour of F2 in this figure is not a continuous one, but piecing together with parallel short bars. At the same time, the spectrogram shows both horizontal and vertical striations spaced at the pitch period or fundamental frequency. They are both due to the voiced excitation. Part (b.) shows the Wigner distribution for this utterance. Compared to (a.) it looks almost as if the vertical scale has been changed, but it is not the case. This representation is dominated by cross-terms that had greater amplitude than the original terms. Part (c.) shows the smoothed Wigner Distribution. In this figure, not only are the contours of formant continuous and protruding, but also the vertical and horizontal bars which correspond to pitch period are effectively removed. Part (d.) gives the schematic spectrogram, extracted contour of formant from (c.) with RD algorithm.

5. CONCLUSION

This paper studies the method of representing Chinese speech with Pseudo Wigner Distribution. Following an introduction to WD and PWD, we discussed our consideration for real implementation of PWD. In addition, an algorithm to extract the formant models from PWD is also described. Experimental results verified the superiority of PWD over conventional STFT.

6. REFERENCE

[1.] Bowen Zhou, “Time-Frequency Representation and Experimental Results Analysis of Chinese Speech Signal”, B.E Thesis, 1996

[2.] Michael D.Railey, “Speech Time-Frequency Representations”, Kluwer Academic Publishers, 1989.

[3.] S.C.Pei and I.I.Yang, “Computing Pseudo-Wigner Distribution by the Fast Hartley Transform”, IEEE Trans. Acoustic, Speech, Signal Processing, Vol. 40 No. 9, pp.2346-2349, Sept, 1992

[4.] J. Jeong，et al , "Kernel design of reduced interference distribution" ， IEEE Trans, Signal processing, vol 40, No 2, 402-412,1992