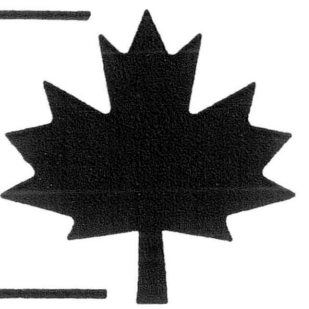
L.E. Atlas, Y. Zhao and R.J. Marks II, "Application of the generalized time-frequency representation to speech signal analysis",
Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing,
pp.517-519, Victoria, B.C. Canada, June 4-5, 1987.

IEEE PACIFIC RIM CONFERENCE ON COMMUNICATIONS, COMPUTERS AND SIGNAL PROCESSING



CONFERENCE PROCEEDINGS

Sponsored by IEEE Victoria Section, IEEE Region 7,
the Faculty of Engineering and the Department of
Electrical Engineering at the University of Victoria

VICTORIA, B.C., CANADA

JUNE 4-5, 1987

IEEE Pacific Rim Conference on Communications, Computers and Signal Processing
June 4-5, 1987

Application of The Generalized Time-Frequency Representation to Speech Signal Analysis

Les E. Atlas
Yunxin Zhao
Robert J. Marks II

Interactive Systems Design Lab
Department of Electrical Engineering, FT-10
University of Washington
Seattle, WA 98195

ABSTRACT

This paper presents a significant result in applying the generalized time-frequency representation (GTFR) to speech signal analysis. A two-dimensional kernel is designed which assures the GTFR of the finite time support property and interference term suppression capability. Comparison of the result with that of the spectrogram and a Wigner distribution shows that much better resolution and accuracy is achieved by using the chosen two-dimensional kernel.

Introduction

The speech spectrogram has been an important tool for speech signal analysis for over 30 years. A spectrogram is a two-dimensional time-frequency representation of speech where the vertical and horizontal axes correspond to frequency and time, respectively. Signal energy is represented by darkness such that formants of speech appear as dark bands on the two dimensional plane [1]. Although the need for such a representation comes from the nonstationarity of signal, the analysis method is based on the assumption of short time stationarity. Therefore, the method cannot give a good result when there are rapid changes of frequency content. This lack of accuracy can occur whenever there are abrupt changes in the manner of articulation, *e.g.* most consonant-vowel boundaries.

Recently, there has been a surge of interest in applying the Wigner distribution (WD) to nonstationary signal analysis and synthesis. The WD is shown in Claasen *et al* [2] to be a time-frequency representation which relates to other time-frequency representations, such as the spectrogram, through a two dimensional kernel. The WD is shown to be the representation which possesses 9 properties that are desirable in signal analysis [2]. Although some encouraging results have been obtained in the application of WD or PWD (a modified version of WD for short time analysis [3,4]), the inherent problem of unavoidable interference terms makes it difficult to use in multi-component signal analysis [3].

It is important to note that an utterance of voiced speech is a multicomponent signal with rich harmonic contents. The PWD of speech has a very complicated time-frequency pattern due to the addition of interference terms on top of the formant frequencies. The consequence is that although attempts have been made in applying the PWD to speech, the spurious interference terms reduce the utility of the time-frequency pattern.

In this paper the problem is approached by side-stepping the need to maintain a WD or PWD. A generalized time-frequency representation is desired such that the kernel of the representation satisfies those constraints which guarantee the desired properties. Significantly better resolution is obtained compared with that of the spectrogram. The mathematical analysis is done mainly on continuous cases for the ease of notation and discrete versions are supplied when it is necessary. All the programming work have been done on a Symbolics 3640 using Zetalisp and the algorithms have been integrated into ISP [5].

Background and Definitions

Wigner distribution

The WD of a real-valued speech signal signal $f(t)$ can be defined both in terms of $f(t)$ and its Fourier transform $F(f)$. In terms of $f(t)$, the definition is

$$W_f(t,f) = \int_{-\infty}^{+\infty} f(t+\tau/2) f(t-\tau/2) e^{-j2\pi f\tau} d\tau$$

Spectrogram (also call short-time Fourier transform)

Let the signal be $f(t)$ and the window be $h(t)$. The segment of the signal selected for short-time analysis is $f_t(\tau) = f(\tau)h(\tau-t)$ and the time-varying spectrogram is

$$S_f(t,f) = \int_{-\infty}^{+\infty}\int f_t(\tau) f_t(\gamma) e^{-j2\pi f(\tau-\gamma)} d\tau\, d\gamma$$

Generalized time-frequency representation

$$C_f(t,f_A;\Phi_M) = \int_{-\infty}^{+\infty}\int\int \Phi_M(f_B,\tau) f(u+\tau/2) f(u-\tau/2) e^{j2\pi(f_B t-\tau f_A-f_B u)} du\, d\tau\, df_B$$

$\Phi_M(f_B,\tau)$ is an arbitrary two-dimensional kernel, the choice of which affects the properties of the time-frequency representation. Two alternative kernel descriptions, which are related to $\Phi_M(f_B,\tau)$ through a one-dimensional forward or inverse Fourier transform, are $\Phi_t(t,\tau)$ and $\Phi_f(f_B,f_A)$. The affect of the choice of $\Phi_t(t,\tau)$, or equivalently $\Phi_f(f_B,f_A)$, on the GTFR is the central issue in the problem of high resolution and distortion-free time-frequency displays.

Kernel Design for High Resolution

An analytic goal in a high resolution display is that of finite time and frequency support. As discussed in [2] two of the nine constraints on the GTFR kernel directly affect these properties. Due to space limitations many results will be stated without proof. Also, the finite time support arguments will be stressed. The interested reader will find more detail in [6].

The finite time support property is most directly affected by the design of $\Phi_t(t,\tau)$. This kernel can be related to the GTFR via

$$C_f(t,f_A;\Phi_t) = \int_{-\infty}^{+\infty}\int \Phi_t(t-u,\tau) f(u+\tau/2) f(u-\tau/2) e^{-j2\pi f_A\tau} du\, d\tau \quad (1)$$

It can thus be seen that the GTFR can be expressed as the Fourier transform with respect to τ of a function $Z(t,\tau)$, where

$$Z(t,\tau) = \Phi_t(t,\tau) * [f(t+\tau/2)f(t-\tau/2)]$$

namely, a convolution with the kernel along the dimension t. It can now be shown graphically that the kernel $\Phi_t(t,\tau)$ chosen in Fig. 1 is sufficient for the property of finite support:

Let the time support of $f(t)$ be $|t| < T$, which is shown in Fig. 2, then the support region of $f(t+\tau/2)f(t-\tau/2)$ on the (t,τ) plane is as shown in Fig. 3. Since the support length adds up through convolution, the support region of $Z(t,\tau)$ will be a rectangle as shown in Fig. 4. Therefore the support region of the GTFR on the (t,f_A) plane is restricted within $|t| < T$, *i.e.* finite time support is maintained. This is shown in Fig. 5.

The kernel $\Phi_t(t,\tau)$ chosen in Fig. 1 can be expressed as

$$\Phi_t(t,\tau) = [1-rect(\frac{\tau}{2|t|})]\, x(t,\tau)$$

where $1-rect(\frac{\tau}{2|t|})$ restricts the support region to the hatched area and $x(t,\tau)$ shapes the kernel within the support region. Since the kernel is designed for short time analysis, the τ dependence of $x(t,\tau)$ should taper $\Phi_t(t,\tau)$ down toward zero at $|\tau| = T$. As with other techniques of spectral analysis, this data taper is needed to reduce the effect of sidelobe leakage.

The choice of t dependence of $x(t,\tau)$ will ultimately relate to the distorting interference terms. By considering $\Phi_f(f_B,f_A)$, which is the the 2-dimensional Fourier transform of $\Phi_t(t,\tau)$, the GTFR can be expressed in an alternative fashion:

$$C_f(t,f_A;\Phi_f) = \int_{-\infty}^{+\infty}\int \Phi_f(f_B,f_A-f')F(f'+f_B/2)F(f'-f_B/2)e^{-j2\pi f_B t}\,df'\,df_B$$

The interpretation of the above GTFR formulation is analogous to Eq. (1). The GTFR can thus be expressed as the inverse Fourier transform with respect to f_B of a function $Y(f_B,f_A)$, where

$$Y(f_B,f_A) = \Phi_f(f_B,f_A) * [f(f_A+f_B/2)f(f_A-f_B/2)]$$

namely, a convolution with the kernel along the dimension f_A. With this formulation, finite frequency support can be graphically obtained for an hourglass-shaped extent of $\Phi_f(f_B,f_A)$, analogous to the argument for finite time support. This choice of $\Phi_f(f_B,f_A)$, is consistent (via a two-dimensional Fourier transform) with the previous choice of $\Phi_t(t,\tau)$.

The reduction of interference terms in the time-frequency display can be brought about by forcing $\Phi_f(f_B,f_A)$ to have low-pass filter behavior in the f_B direction, thus smoothing out the spurious components. Ideally, this corresponds to spreading $\Phi_t(t,\tau)$ as much as possible in the support region in the t direction. If $x(t,\tau)$ is chosen to be independent of t and Gaussian in the τ direction, *i.e.*

$$x(t,\tau) = e^{-2\alpha\tau^2}$$

the design constraints are satisfied.

Discrete Case

The discrete version of the GTFR can be defined as

$$C_f(n,\theta;\Phi_n) = \sum_{p=n-L}^{n+L}\ \sum_{k=-L}^{L} \Phi_n(n-p,k)f(p+k)f(p-k)e^{-j2k\theta}$$

where $2L-1$ is the number of data points in the analysis segment. The discrete version of the appropriate kernel is

$$\Phi_n(n,k) = rect(\frac{n}{|k|})e^{-2\alpha k^2}$$

where the $rect(\,.\,)$ function is defined as being unity for arguments between -1 and 1, inclusive.

Experimental Results

The sentence "that you may see" and the single word "jump" were used in the experiments. The passages were low-pass filtered at 5 kHz and sampled at 20 kHz. The duration of the waveforms are 1.9712 sec. and 0.1536 sec., respectively. All displays are made in a spectrogram-like fashion: the horizontal axis corresponds to time, the vertical axis corresponds to increasing frequency, and increasing magnitude of the time-frequency display is indicated by increasing darkness. (Negative values for the GTFR displays were set to zero.) Also, all displays have a frequency range of 5 kHz.

In Fig. 6, the top pane is the speech "that you may see", the middle one is the spectrogram and the bottom one is the GTFR. The spectrogram has a DFT size of 256 samples and the analysis interval of 32 samples. For the GTFR the DFT size is 127 samples and the analysis interval is 3 samples. The comparison of the middle and the bottom panes shows more precisely defined formant tracks in the GTFR. The most prominent part is in the coarticulation between "that" and "you", where the formant tracks which link those of "that" and "you" are clearly visible in the GTFR but are almost smeared out in the spectrogram.

Fig. 7 shows the spectrogram, the PWD and the GTFR of the speech word "jump" on the top, middle and bottom pane, respectively. The spectrogram and the PWD are the same as in Fig. 6. For the GTFR the parameters of DFT size and analysis interval are 127 samples and 8 samples, respectively. The comparison of the three panes shows that the GTFR gives much clearer formant tracks in the "j" sound than the spectrogram and the PWD do. Due to the shorter analysis interval the vertical striations can now be observed in the GTFR, in contrast to the GTFR in Fig. 6.

Fig. 8 shows the result of narrow band analysis on the sentence "that you may see". The spectrogram is on the top pane with a DFT size of 256 samples and an analysis interval of 64 samples. The GTFR is on the bottom has a DFT size of 128 samples and an analysis interval of 64 samples. The GTFR has much higher resolution of formant frequency positions.

Fig. 9 shows the case when the word "jump" is contaminated by additive white noise with the signal to noise ratio being -10 db. From the top to bottom pane are the speech waveform, the spectrogram, the PWD and the GTFR. Due to a pre-emphases operation for all 3 time-frequency displays, the noise has more effect on the high frequency content than it does on the low frequency part. In this case, the PWD fails completely since no formants can be observed. The spectrogram still shows some formant structure in the low frequency part but is not as clearly discernible as in the GTFR.

Summary

A GTFR with finite time support property and enough interference term suppression capability is obtained by imposing constraints on the two dimensional kernel design. Experiments show that much clearer formant tracks can be obtained using the GTFR than by using the conventional spectrogram. This nonparametric method with its precisely defined formant locations might provide an alternative to LPC for speech recognition, especially in the presence of noise. Although the GTFR has been designed primarily for the purpose of speech analysis, there is no doubt that it can be applied to other time series analysis as well. By increasing the dimensionality of the kernel and by trading temporal for spatial variables it may also be possible to extend the

application to the field of image processing. Moreover, this technique of designing kernels according to a set of constraints on two-dimensional planes will yield GTFRs more satisfactory than existing application-specific transforms.

Acknowledgements

This research was supported by a National Science Foundation Presidential Young Investigator Award and by a contract from Boeing Computer Services. The authors would also like to thank Richard Lyon of Schlumberger Palo Alto Research for providing much of the software used in the experimental studies.

References

1. L.R.Rabiner and R.W.Schafer, **Digital Processing of Speech Signals**, Prentice-Hall, Englewood Cliffs, NJ, 1978.
2. T.A.C.M. Claasen and W.F.G. Mecklenbrauker, "The Wigner Distribution, A Tool for Time-frequency Signal Analysis. Part 3: Relations with Other Time-frequency Signal Transformations," *Philips J. Res.* **35**, pp. 373-389, 1980.
3. P.Flandrin, "Some Features of Time-frequency Representations of Multicomponent Signals," *Proc. ICASSP*, San Diego, CA, Mar. 1984, pp. 41B.4.1-41B.4.4.
4. T.A.C.M. Claasen and W.F.G. Mecklenbrauker, "On the Time-frequency Discrimination of Energy Distributions: Can They Look Sharper than Heisenberg?" *Proc. ICASSP*, San Diego, CA, Mar. 1984, pp. 41B.4.1-41B.4.4.
5. G.Kopec, "The Integrated Signal Processing System ISP," *IEEE Trans. Acoust., Speech, Sig. Proc.* **ASSP-32**, Aug. 1984.
6. Y.Zhao, L.Atlas, and R. Marks, "Speech Signal Analysis with a Generalized Time-Frequency Representation." (Submitted for publication to *IEEE Trans. Acoust., Speech, Sig. Proc.*)

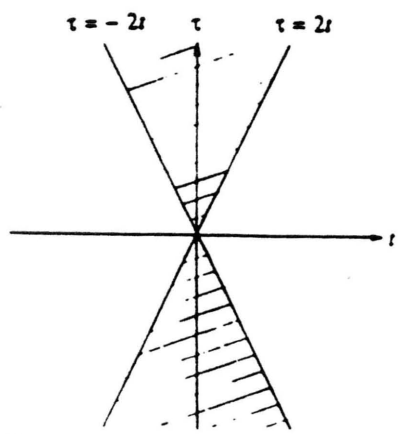


Fig. 1: The support region of kernel $\Phi_t(t,\tau)$

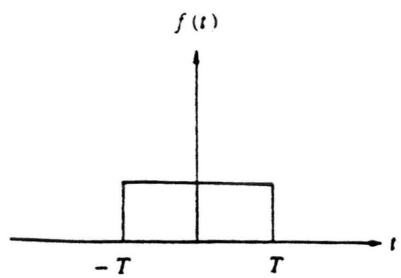


Fig. 2: The support of $f(t)$.

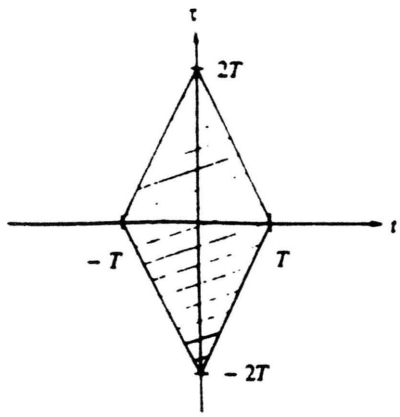


Fig. 3: The support region on the (t,τ) plane of $f(t+\tau/2)f(t-\tau/2)$.

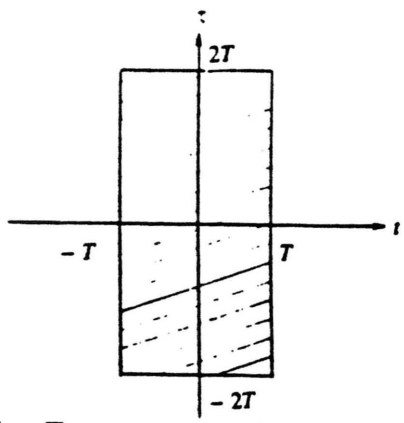


Fig. 4: The support region on (t,τ) plane of $\Phi_t(t,\tau) * [f(t+\tau/2)f(t-\tau/2)]$

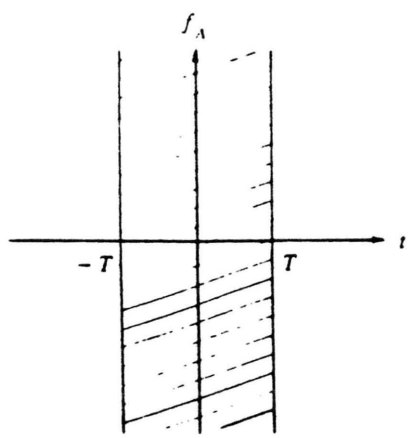


Fig. 5: The support region on $(t\,f_A)$ plane of the GTFR.

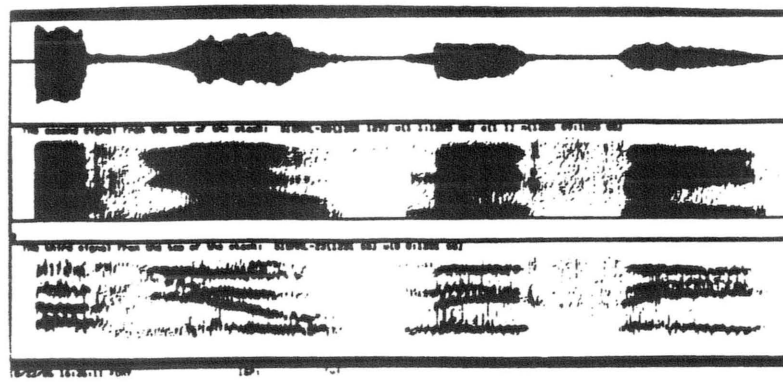


Fig. 6: A comparison of the spectrogram and the GTFR on the sentence "that you may see".

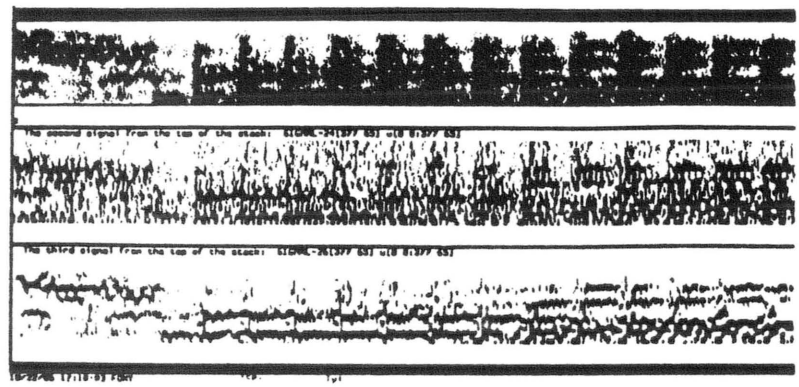


Fig. 7: A comparison of the spectrogram, the PWD and the GTFR on the word "jump".

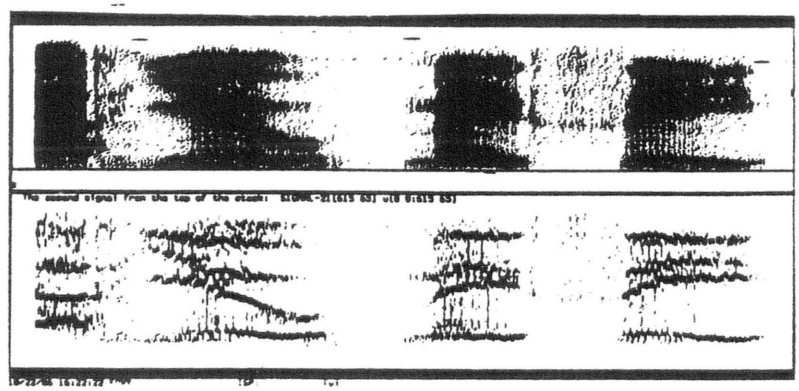


Fig. 8: A comparison of the spectrogram and the GTFR on the sentence "that you may see" using narrow band analysis.

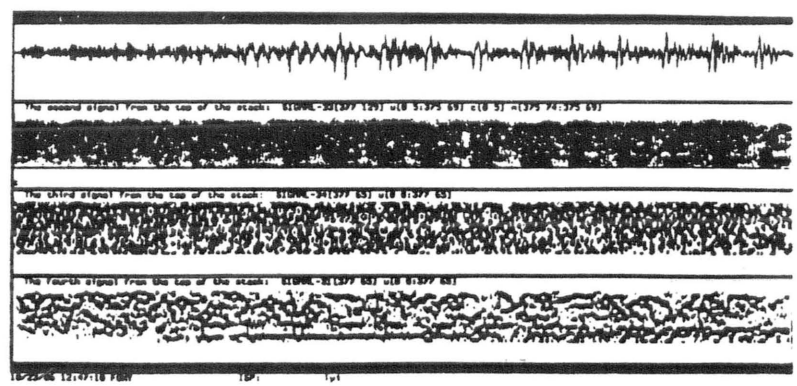


Fig. 9: A comparison of the spectrogram, the PWD and the GTFR on the word "jump" in the presence of noise.