New Features Based on the Cohen's Class of Bilinear Time-Frequency Representations for Speech Recognition

Jingdong Chen   Bo Xu and Taiyi Huang

National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, P. O. Box 2728, Beijing, China

e-mail: cjd@prldec3.ia.ac.cn

Abstract: Although short-time Fourier Analysis based features such as LPCC and MFCC have been widely used in state-of-the-art speech recognizers, the short-time analysis technique suffers from the well-known trade-off between time and frequency resolution and works under the assumption that speech signal is short-time stationary. This paper investigates an approach to use Cohen's class bilinear time-frequency distributions representing speech signal for speech recognition. Preliminary experiments show that the new feature can better represent speech signal and can improve the accuracy of a speech recognizer.

1. INTRODUCTION

Time-frequency (TF) representations have been used extensively for speech analysis and speech recognition. State-of-the-art systems for speech recognition segment the speech signal into short intervals on the order of tens of milliseconds. Short-time analysis is then used to estimate the parameters of each segment under the implicit assumption that the signal is quasi-stationary over the intervals. However, short-time analysis, so far the most common tool for analyzing the speech signals, suffers from the well-known trade-off between time and frequency resolution. If one specifies the instant at which the frequencies occur, one needs to shorten the time window. However, if one decreases the time window, namely, if one locates the events in time, the frequency resolution is reduced. In addition, too short a window produces a poor spectral representation of the speech, is sensitive to temporal effects of voicing, and performs poorly in noise. While a longer window alleviates these effect somewhat, the quasi-stationary assumption is often violated, particularly for the segment which transits from consonant to vowel. Hence, these fast time-varying segments of speech are often misclassified.

Recently, there has been a surge of interest in the application of TF distributions to speech processing [10][11], which seeks a way to accurately represent the energy of a signal jointly in time and frequency. The Wigner distribution (WD) is the first TF distribution. In 1966, a generalized time-frequency representation named Cohen's class of time-frequency distribution is proposed by L. Cohen[1]. Cohen's class TF distribution is a kind of bilinear transform which makes no implicit short-time stationary assumption and can get very high time and frequency resolution simultaneously.

This paper investigates a new technique which uses Cohen's class bilinear TF distribution to represent speech signal in the front-end processing of speech recognition. Experimental results show that, comparing with the LPCC and MFCC, the new technique can improve the accuracy of a speech recognizer.

The remainder of this paper is arranged as follows: In section 2, the definition and some kernels of Cohen's class TF representation are introduced. And a new kind of feature based on the Cohen's class TF representation is described in section 3. Section 4 presents some experimental results on a speaker-dependent speech recognizer. And some important conclusions are presented in section 5.

2. Cohen's class TF distribution

Cohen's class of distribution, by definition, is

$$P(t,\omega)=\frac{1}{4\pi^2}\iiint\phi(\theta,\tau)f(u+\frac{\tau}{2})f^*(u-\frac{\tau}{2})\cdot e^{j\theta u-j\theta t-j\omega\tau}dud\tau d\theta \quad (1)$$

Where $f(u)$ is the time signal, $f^*(u)$ its complex conjugate and $\phi(\theta,\tau)$ the kernel defining a particular distribution. This process can be seen more clearly by rewritten the above formula as[2]

$$P(t,\omega)=\frac{1}{4\pi^2}\iint A(\theta,\tau)\phi(\theta,\tau))e^{-j\theta t-j\omega\tau}d\tau d\theta \quad (2)$$

where $A(\theta,\tau)$ is called the symmetrical AF and is given by

$$A(\theta,\tau)=\int f(t+\frac{\tau}{2})f^*(t-\frac{\tau}{2})e^{j\theta t}dt \quad (3)$$

(2) and (3) indicate that Cohen's class of distribution can be interpreted as the 2-D Fourier Transform of a weighted version of the ambiguity function (AF) of the signal to be analyzed.

The kernel $\phi(\theta,\tau)$ may depend on the signal $f(t)$, but if it is independent of the signal, the TFD is said to have a bilinear structure of the signal. Different choices for the kernel function $\phi(\theta,\tau)$ yields different TFD's. For example, if a kernel is taking a constant value, say 1, i. e.

$$\phi(\theta,\tau)=1 \quad (4)$$

With this choice of kernel, a well-known TFD called Wigner distribution (WD) is yielded. The WD satisfies a long list of desirable properties [3], such as it meets marginal conditions, i. e.

$$\text{P1.} \quad \int_{-\infty}^{\infty} P(t,\omega)d\omega=|s(t)|^2 \quad (5)$$

$$\text{P2.} \quad \int_{-\infty}^{\infty} P(t,\omega)dt=|s(\omega)|^2 \quad (6)$$

The WD has the ability to provide simultaneous high resolutions in time and frequency axes which exceeds that of the short time Fourier method, thus avoid the TF resolution tradeoff that post on the short-time analysis technique. However, When the WD is used to analyze speech signal or some other kinds of multicomponent signals, it produces cross-terms between two frequency components at different locations in the time-frequency plane, which are difficult to interpret. To suppress cross-terms, many other kinds of kernels are proposed, such as Choi-williams kernel[4].

$$\phi(\theta,\tau)=e^{-\theta^2\tau^2/\sigma} \quad (7)$$

Where σ is a free parameter that is used to control the amounts of cross-terms. If this parameter is chosen appropriately, the cross-terms may reduce to some extent, but it can not eliminate the problem entirely.

In addition to the time and frequency marginal conditions and cross-terms, it is desired that the TFR meets some other desired properties. Some of these properties are:

P3. Energy conservation

P4. Real-valued

P5. Translation covariance

P6. Dilation covariance

P7. Wide-sense support conservation

P8. Instantaneous frequency concentration

P9. Perfect localization on linear chirp signals

P10. Nonnegative

About the details, readers may refer to [1]. Many works have shown that a Cohen's class distribution with bilinear structure does not simultaneously satisfy the properties described above. For example, if it meets the property of nonnegative (P10), it will not satisfy the two marginal conditions (P1, P2), which means that the Cohen's class bilinear distribution are not true distributions. But we still use "distribution" as commonly used in the literature.

In this paper, besides the Wigner and Choi-williams distributions, several other distributions are also selected to provide better representation of speech signal and improve the accuracy of the speech recognition system. They are:

1. Margenau-Hill distribution (MH) [5]

$$\phi(\theta,\tau)=\cos(\pi\theta\tau) \quad (8)$$

2. Compound distribution [6]

$$\phi(\theta,\tau)=\exp(-2\pi^2\theta^2\tau^2/\sigma^2)\cos(2\pi\beta\theta\tau) \quad (9)$$

where σ and β are two free parameters used to control the cross-terms of the distribution.

3. Bessel distribution [7]

$$\phi(\theta,\tau)=\frac{J_1(2\pi\alpha\theta\tau)}{2\pi\alpha\theta\tau} \quad (10)$$

Where $J_1(\cdot)$ is the first kind Bessel function of order one, α is a positive scaling factor.

4. XY distribution [8]

$$\phi(\theta,\tau)=\frac{g(\theta,\tau)}{1+\alpha\theta^2\tau^2} \quad (11)$$

Where $g(\theta,\tau)$ is a radially non-increasing function and α is a free parameter which controls the spread of the kernel function.

5. Gaussian distribution [8]

$$\phi(\theta,\tau)=g(\theta,\tau)[e^{\alpha n\theta^2}+e^{-\beta n\tau^2}]^{1/n} \quad (12)$$

Where $g(\theta,\tau)$ is also a radially non-increasing function and α and β are free parameters which control the spread of the kernel function, and n controls the flatness of the kernel along the axes.

3. Temporal cepstrum

For the use of Cohen's class of Bilinear TF distribution in speech recognition, a new feature called temporal cepstrum is introduced in this section. The procedure of the new feature is shown as follows.

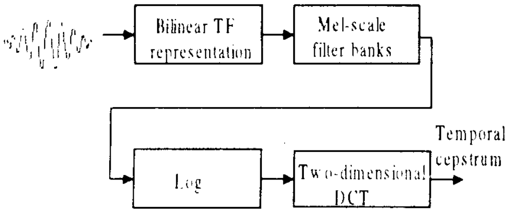

Fig. 1 Temporal cepstrum

For input speech signal, we first estimated its Bilinear TF distribution by using the kernels described above. And for each time of interest, a mel-scale triangle filterbank is used to smooth a slice of the TFD estimated. The log operation is then applied to the magnitude of the TFD, which has been revealed to have at least two effects, one is for compressing the dynamic range of the TFD, and another is to change the multiplicative components in the spectral domain into additive ones in the log-spectral domain. And finally, several slices of TFD combined together, a two dimensional discrete cosine transform (TDCT) is used to decorrelate the log-TFD to allow the subsequent statistical model to use diagonal matrix, and it also has the effect of compressing the log-TFD into lower-order coefficients.

4. experiments

The experiments have been performed to evaluate the recognition performance by using the new acoustic parameter. For comparison, the recognition results based on the MFCC are also presented.

The database used is spoken in mandarin. It consists of 190 isolated words spoken by five male speakers (m1,m2,m3,m4,m5) and one female speaker (f1). For each speaker, each word is uttered seven times. The database is originally recorded through a telephone switch and is sampled at 8KHz. For the utterances of each word, the first five ones are recorded with a telephone channel, and another two utterances are sampled with two other different telephone channels. In our experiments, the first five utterances of each word are used for training the HMM models, and another two utterances are for testing.

The speech recognizer is a speaker-dependent one which is based on continuous density HMM using whole word models. Models are left-to-right with no state transitions. Eight states are used for each model. The training iteration begins with uniformly probabilistic models.

To estimate the MFCC, the speech is segmented into frames of 48ms (384 samples) with 24ms overlap. The data is preprocessed using a 48ms Hamming window and preemphasized with a factor of 0.97. In temporal cepstrum estimation, we still segment the speech signal into frames of 48ms but with 42ms overlap. A slice of TFD is estimated every 6ms. And Four slices of TFD combined together, we use a TDCT to decorrelate the log-TFD into 12 cepstral coefficients.

The experimental results are shown in Table I, Table II, Table III and Table IV respectively.

Baseline	Word error rate
MFCC	11.6%

Table I. The baseline performance of the recognizer (Speaker is m1)

Distribution	Word error rate	Error reduction
Wigner	11.1%	4.3%
Choi-Williams	10.0%	13.8%
MH	10.5%	9.9%
Compound	8.9%	23.3%
Bessel	8.4%	27.6%
XY	8.4%	27.6%
Gaussian	8.9%	23.3%

Table II. The performance for different kernels (Speaker is m1)

Baseline	Word error rate
MFCC	8.4%

Table III. The baseline performance of the recognizer (The speaker is f1)

Distribution	Word error rate	Error reduction
Wigner	7.9%	6.0%
Choi-Williams	6.8%	19.0%
MH	6.8%	19.0%
Compound	6.3%	25.0%
Bessel	5.8%	31.0%
XY	5.3%	36.9%
Gaussian	5.8%	31.0%

Table IV. The performance for different kernels (The speaker is f1)

From the results, it can been seen that the use of Cohen's class of bilinear TF distribution can improve the accuracy of the speech recognizer. However, for different kernels, the improvements are different. Up-to-date, we have not found theoretical rules that can be used to determine which kernel is better. Perhaps, this is one of the most important concentration in our recent future work.

For each kernel, if the free parameter is changed, the recognition performance may be different. For Choi-williams kernel, the free parameter is tested with a value of 1.0, 5.0, and 10.0 respectively, the recognition rates are about the same. The result presented in the Table II and Table IV are for $\sigma = 1$. For compound kernel, the σ is tested with a value of 0.5, 1.0, 2.0and 5.0, the β is tested with a value of 1/3, 1/2 and 1 respectively, we find that when $\sigma = 2.0$ and $\beta = 1/2$, the result is the best one which is shown in the Tables above. The Bessel kernel is tested with α of a value of 0.4, 0.45 and 0.5, the difference of α makes very little difference in the error rate. For the XY kernel, the free parameter α is tested with a value of 0.5, 0.6, 0.7, and 0.8. The $g(\theta, \tau)$ is taking a constant value, say 1. The results shown in Table II and Table IV are for $\alpha = 0.7$. These results are significantly better than that for $\alpha = 0.5$, $\alpha = 0.6$ and $\alpha = 0.8$. For Gaussian kernel, the $g(\theta, \tau)$ is also taking a value of 1 for all θ and τ. The free parameter α is tested with a value of 10^4, 10^5 and 10^6, and β with 10^4, 10^5 and 10^6. n is set to 1. The results shown are for $\alpha = 10^5$ and $\beta = 10^5$.

5. Conclusion

This paper introduces a new feature of speech signal called temporal cepstrum which is based on the Cohen's class bilinear time-frequency distribution. Experiments show that, comparing with the LPCC and MFCC which are based on the traditional short-time Fourier analysis technique, the temporal cepstrum can improve the recognition accuracy. However, for different kernels, the improvements are different. So far, there are no mathematical rules for determining which kernel is better for speech recognition. Our further work may concentrate on the following two aspect, one is to seek kernels which can accurately represent the speech signal, another is to find more effective numerical implementation algorithms for the Cohen's class of Time-frequency distribution.

Reference

[1] Leon Cohen, "Time-Frequency Distributions — A Review", Proceedings of the IEEE, Vol. 77, No. 7, July, 1989.

[2] Douglas L. Jones and Richard G. Baraniuk, "An Adaptive Optimal-Kernel Time-Frequency Representation," IEEE Transactions on Signal Processing, Vol. 43, No. 10, October 1995.

[3] W. Martin and P. Flandrin, "Wigner-Ville Spectral Analysis of Nonststionary Processes", IEEE Transactions on Acoustics , Speech and Signal Processing, Vol. ASSP-33, No. 6, December 1985.

[4] H. I. Choi and W. Williams, "Improved Time-Frequency Representation of Multicomponent Signals Using Exponential Kernels," IEEE Transactions on Acoustic, Speech and Signal Processing, Vol. 37, June 1989.

[5] H. Margenau and R. N. Hill, "Correlation Between Measurements in Quantum Theory," Prog. Theor. Phys., Vol. 26, 1961, PP. 722-738.

[6] Bilin Zhang and Shunsuke, "A Time-Frequency Distribution of Cohen's Class with a Compound Kernel and its Application to Speech Signal Processing", IEEE Transaction on Signal Processing, Vol. 42, No. 1, January 1994.

[7] Zhenyu Guo, Louise-Gilles and Howard C. Lee, "The Time-Frequency Distributions of Nonstationary Signals Based on a Bessel Kernel," IEEE Transactions on Signal Processing, Vol. 42, No. 7, July 1994, PP. 1700-1707.

[8] Adam B. Fineberg and Kevin C. Yu, "Time-Frequency Representation Based Cepstral Processing for Speech Recognition," In Proceeding of ICASSP'96, Atlanta, Georgia, USA, May 1996, PP. 25-28.

[9] Richard N. Czerwinski and Douglas L. Jones, "Adaptive Cone-Kernel Time-Frequency Analysis," IEEE Transaction on Signal Processing, Vol. 43, No. 7, July 1995.

[10] James W. Pitton, Les E. Atlas and Patrick J. Loughlin, "Application of Positive Time-Frequency Distributions to Speech Processing," IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 4, October 1994.

[11] James W. Pitton, Kuansan Wang and Biing-Hwang Juang, "Time-Frequency Analysis and Auditory Modeling for Automatic Recognition of Speech", Proceedings of the IEEE, Vol. 84, No. 9, Spetember 1996.