# CONSONANT RECOGNITION USING THE COVARIANCE OF THE PSEUDO WIGNER DISTRIBUTION

D.Rainton and S.J Young
Cambridge University Engineering Department, U.K.

## ABSTRACT

It is generally accepted that the consonant in a consonant-vowel(CV) pair can be identified by the nature of the formant transitions in the vowel. STFT power spectral *snapshots* fail to capture the detailed time-varying nature of these transitions. In this paper we show that such spectra can be considered weighted time averages of the pseudo-Wigner distribution (PWD) when appropriate gaussian windows are used in the computation of both. Given this interpretation we then speculate as to whether the higher order statistics of the PWD convey additional consonant discriminant information. Experimental evidence indicates that they do.

## INTRODUCTION

The formant transitions in a CV syllable are known to contain useful consonant discriminant information[Liberman 1956]. From a human perceptual point of view however, authors disagree as to these transitions importance relative to the noise burst at stop release. Stevens and Blumstein[Stevens & Blumstein 1978] for example, argue that invariant cues for place of articulation of stop consonants can be characterised independently of the following vowel. In their view, then, formant transitions are *not* the primary cue signaling place of articulation. However for the engineer attempting to build practical speech recognition devices, extracting maximum information about the consonant from the neighbouring formant transitions promises to provide more robust performance in noisy environments. Typically formant transitions contain more signal energy than fricative bursts and so are less susceptible to noise corruption. The problem is to find a suitable choice of feature space which captures the nonstationary nature of these transitional regions.

Most of todays commercial speech recognition systems rely on a short time Fourier power spectral representation ($S_x(t,\omega)$) of the input signal, typically sampled at around 100 Hz. $S_x(t,\omega)$ is an easily interpreted, non negative time-frequency decomposition, defined by the equation

$$S_x(t,\omega) = \left| \int_{-\infty}^{+\infty} h(t-\tau)x(\tau)e^{-j\omega\tau}d\tau \right|^2 . \tag{1}$$

The recognition of *steady state* vowels is considered very feasible using this quasi-stationary estimator. Most researchers would agree that in this instance the STFT feature space contains the important information, exhibited in an important way. However, such a transformation is less adequate for characterising rapidly changing events such as the transition into or out of a stop consonant. This is because the ability of this transformation to localise energy in the time-frequency plane is strictly limited by the resolution of the window $h(t)$. An increase in time resolution results in a decrease in frequency resolution and vica-versa[Priestley 1981]. Furthermore, imposing the condition of stationarity upon data which is inherently non-stationary can result in an inconsistent or misleading spectral representation[Silverman & Lee 1987].

## THE WIGNER DISTRIBUTION

A detailed examination of formant transition characteristics requires a spectral representation with a greater time-frequency resolution than that provided by $S_x(t,\omega)$. In recent years an alternative time-frequency representation known as the Wigner distribution (WD) has received growing interest[Lowe, Tomlinson & Moore 1986]. The WD of a signal $x(t)$ is defined as

$$W_x(t,\omega) = \int_{-\infty}^{+\infty} x(t+\frac{\tau}{2})x^*(t-\frac{\tau}{2})e^{-j\omega\tau}d\tau \tag{2}$$

where $*$ denotes complex conjugate. In practice the evaluation of the WD at $t = t_0$ is achieved by sampling and then multiplying the signal $x(t)$ by a real symmetric window $h(t)$ centered at $t_0$. As the WD is simply the Fourier transform of the product of the signal and its time inverted conjugate the usual window carpentry can be used in choosing $h(t)$. This modified WD is known as the discrete pseudo-Wigner distribution(PWD) and is defined by the equation

$$\widetilde{W}_x(t,\omega_n) = 2\sum_{k=-N+1}^{N-1} |h_N(k)|^2 x(t+k)x^*(t-k)e^{-j2k\pi(n/N)} \qquad \omega_n = \pi(n/N). \tag{3}$$

In this paper discrete time versions of the various spectral estimators are indicated by a subscript on the frequency parameter. For computational purposes (3) is often rearranged in the form

$$\widetilde{W}_x(t,\omega_n) = 4Re\left\{\sum_{k=0}^{N-1} |h_N(k)|^2 x(t+k)x^*(t-k)e^{-j2k\pi(n/N)}\right\} - 2|x(t)|^2 \tag{4}$$

Choosing $N$ to be a power of 2 allows efficient calculation of the PWD over $2N-1$ points to be achieved using an ordinary radix-2 fft over $N$ points. An unfortunate consequence of the PWD as defined in (3) is the occurrence of aliasing even when the sampling of $x(t)$ satisfies the Nyquist criterion. Several authors have discussed ways of overcoming this problem[Claasen & Mecklenbraüker 1983]. The solution chosen in this paper is to transform the signal into its analytic equivalent prior to computing the PWD. This has the added advantage of preventing interference between positive and negative frequency components in the resulting distribution.

## SOME ADVANTAGES AND DISADVANTAGES OF THE PWD

The properties of the WD and PWD have been exhaustively documented elsewhere[Claasen & Mecklenbraüker 1980]. Here we simply point out several important properties which are relevant to the current discussion.

The main advantage of the PWD is the absence of smoothing in the time domain. Windowing the WD in time produces smearing in the frequency direction only, hence, the frequency resolution of the PWD can be increased by using a longer window without loss of time resolution. Stated more formally

$$\widetilde{W}_x(t,\omega_n) = \frac{1}{2\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} W_x(t,\eta)W_h(0,\omega_n-\eta)d\eta \tag{5}$$

where $W_h(t,\omega_n)$ is the discrete WD of the window $h(t)$. The disadvantages of the PWD (and WD) are its lack of positivity and its bilineal nature. The first prevents any local interpretation of an energy density in the time-frequency plane. The second gives rise to interference or cross terms[Hlawatsch 1984] which may cause difficulties in interpretation, especially in the case of complex multicomponent signals such as speech. A further disadvantage of the PWD is the very high data rates that it generates. Because of the absence of smoothing in the time domain, the PWD must be sampled at a similar frequency to the signal itself. Failure to do so can result in visual aliasing when viewing an undersampled distribution.

In most practical situations[Allard 1988], it is necessary to apply some sort of smoothing to the WD in the time domain as well the frequency domain in order to suppress negative regions and cross terms. This can be achieved by simply low pass filtering the PWD in time. The smoothed pseudo-Wigner distribution[Allard 1988] (SPWD) is related to the PWD by the equation

$$\widetilde{\widetilde{W}}_x(t,\omega_n) = g(t) * \widetilde{W}_x(t,\omega_n) \tag{6}$$

where $g(t)$ is a real symmetric window. This smoothing can be incorporated directly into the calculation of the PWD by rewriting (4) as

$$\tilde{\tilde{W}}_x(t,\omega_n) = 4Re\Big\{\sum_{k=0}^{N-1} |h_N(k)|^2 \sum_{l=-L+1}^{L-1} g(l)x(t+k+l)x^*(t-k+l)e^{-j2k\pi(n/N)}\Big\} - 2\sum_{l=-L+1}^{L-1} g(l)|x(t+l)|^2. \tag{7}$$

In the next section we show that $\tilde{\tilde{W}}_x(t,\omega_n)$ computed using an appropriate choice of $g(t)$ and $h(t)$ is equivalent to the short time Fourier power spectrum.

## THE RELATIONSHIP BETWEEN THE PWD AND THE FOURIER SPECTROGRAM

The spectrogram is related to the WD through a 2-d convolution[Claasen & Mecklenbraüker 1980]

$$S_x(t,\omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} W_x(\tau,\eta)W_h(\tau - t, \omega - \eta)d\tau d\eta \tag{8}$$

where $W_h(t,\omega)$ is the WD of the window $h(t)$ defined as

$$W_h(t,\omega) = \int_{-\infty}^{+\infty} h(t+\frac{\tau}{2})h^*(t-\frac{\tau}{2})e^{-j\omega\tau}d\tau. \tag{9}$$

If we select $h(t)$ to be a Gaussian window of the form

$$h(t) = (2\pi)^{-\frac{1}{4}}\sigma^{-\frac{1}{2}}exp(\frac{-t^2}{4\sigma_t^2}) \tag{10}$$

then we obtain from (9) the double Gaussian kernel

$$W_h(t,\omega) = \frac{1}{\sigma_t\sigma_w}exp(-\frac{1}{2}[\frac{t^2}{\sigma_t^2} + \frac{\omega^2}{\sigma_w^2}]) \tag{11}$$

where $\sigma_t\sigma_w = \frac{1}{2}$.

Thus the spectrogram $S_x(t,\omega)$ computed using the Gaussian window $h(t)$ defined in (10) is identical to the WD smoothed with the double Gaussian kernel defined in (11). Given that $W_h(t,\omega)$ is in this instance a separable function we can rewrite (8) in the form

$$S_x(t,\omega) = [W_x(t,\omega) * K(\omega)] * k(t) \tag{12}$$

where

$$K(w) = \frac{1}{\sigma_w}exp(-\frac{1}{2}w^2/\sigma_w^2) \tag{13}$$

and

$$k(t) = \frac{1}{\sigma_t}exp(-\frac{1}{2}t^2/\sigma_t^2) \tag{14}$$

The bracketed term in (12) is simply a pseudo-Wigner distribution of $x(t)$. Thus $S_x(t_0,\omega)$ computed using the Gaussian window h(t) is proportional to a weighted time average of the PWD about $t = t_0$ where the weighting function is also Gaussian

$$S_x(t,\omega) = \tilde{W}(t,\omega) * k(t). \tag{15}$$

In this the rest of this paper when we refer to the relationship between the short time Fourier power spectrum and the PWD we implicitly assume that both have been computed so as to satisfy the above conditions.

If we characterise a frame of speech by a single STFT power spectral vector instead of its PWD then we effectively loose information about the higher order statistics of the PWD. The significance of this information loss for the accurate characterisation of non stationary signals such as formant transitions is discussed in the next section.

## CHARACTERISING FORMANT TRANSITIONS USING THE PWD AND STFT

Each CV transition region can be represented by a distribution of PW vectors in a frequency space where each vector represents the PWD at time $t$, the distribution having being computed at the signal sample rate. Such a representation involves no information loss as the origin signal can be trivially recovered from the distribution. Obviously however the distribution does contain a high degree of redundancy. The problem of distinguishing between different CV transitions can effectively be translated into one of distinguishing between different PW distributions. How then do we best characterise such distributions for classification purposes?

Representing each transition region by one or more STFT power spectral vectors is one popular alternative. As pointed out in the last section, characterising a distribution of PW vectors by a single STFT power spectral vector computed over the same time interval is equivalent to representing the PWD by its time weighted mean. Classification of the PW distributions based on such information would clearly be optimal if the distributions all had the same shape and spread, differing only in their average positions in the frequency space.

Unfortunately, this turns out not to be the case. Transitions from different consonants into the same vowel have PW distributions with very different shapes and spreads. Discrimination based on a single STFT power spectral vector will ignore such differences which are related to the higher order statistics of the PWD. In practice however, the transition region is usually of sufficient duration to be represented by a sequence of several STFT power spectral vectors. This is equivalent to representing the entire PWD by several localised averages. This will convey some additional information about the shape and spread of the PWD but introduces time alignment problems as multiple vectors now represent a single speech event.

Here we pursue a different approach to characterising the PWD. Instead of simply looking at the first order statistics of the PWD as is the case with the gaussian STFT spectral estimator one obvious way to make use of the additional information available in the PWD is to examine its higher order statistics. In the experiment described in the rest of this paper we do just that and look at the discriminant information contained within the covariance of the PWD.

## CALCULATING THE COVARIANCE OF THE PWD

In order to compute the covariance matrix from the PWD we define a vector $X_t$ representing the PWD at time $t$ by the equation

$$X_t = (x_{t1}, x_{t2}, ..., x_{tN})^T \qquad \text{where} \qquad x_{tk} = \widetilde{W}_x(t, w_k) \qquad \text{k=1,...,N} \;. \tag{16}$$

An $N$ by $N$ sample covariance matrix $\Sigma$ is constructed whose elements $\Sigma_{ij}$ are defined by the equation

$$\Sigma_{ij} = \frac{1}{T}\sum_{t=0}^{T}[(x_{ti} - \overline{x}_i)(x_{tj} - \overline{x}_j)] \tag{17}$$

where

$$\overline{x}_l = \frac{1}{T+1}\sum_{t=0}^{T} x_{tl} \tag{18}$$

To quantise the difference between different $\Sigma$ matrices, the covariance matrix is treated as a space vector with $N(N+1)/2$ dimensions. A Euclidian Distance measure is then used to compute the distance between these covariance *vectors*.

## EXPERIMENT

### Purpose

This experiment was designed to discover whether the PW covariance matrix $\Sigma$ defined by (17) can be used to distinguish between the different consonants in a set of CV syllables where the vowel portion of each syllable is the same. Only the transition regions in each waveform were made available to the spectral estimators. Classification was then achieved using simple nearest neighbour (nn) classifiers. A comparative study of nn recognition performance using both the PW covariance and the more traditional Fourier and lpc short time power spectral estimators was conducted.

Data

The data for this experiment was produced by a single male speaker with a Southern British accent speaking each of the eight syllables *ba, da, ga, pa, ta, ka, ma, na* ten times. This data set was chosen to provide a set of consonants varying in both place and manner of articulation, including voiced stops (*ba, da, ga*), unvoiced stops (*pa, ta, ka*) and the nasals (*ma, na*).

The recorded speech was sampled at 10kHz and digitised with 12bit resolution. The region of formant transition in each syllable was then manually extracted using the Cambridge interactive speech editor (CAMSED), coupled to a spectrogam display system. This provided 80 speech files each containing a single CV transition, typically around 60ms in length. A spectrographic analysis using pseudo-Wigner and short time Fourier spectral estimators was then performed on each file. The short time Fourier power spectrum was computed at intervals of 5ms using a 12.8ms gaussian window. The pseudo Wigner spectrum was computed at each sample point using a 12.8ms gaussian window and a single PW covariance matrix generated for each file. In addition we also performed a 12th order lpc autocorrelation analysis on each speech file. This used a 12.8ms hamming window with a 7.8ms overlap. Thus each file provided a sequence of STFT power spectral vectors, a sequence of lpc coefficient vectors and a single PW covariance *vector.*

Classification

A k nearest-neighbour classification scheme (k=1) was chosen with dynamic time warping for time alignment of the STFT and lpc coefficient vector sequences. Three separate classifiers were built, one for each data type. Euclidian distance measures were used in the STFT and PW covariance recognisers and an Itakura distance in the lpc recogniser. The need for data to train the classifiers and additional data to test them presented a familiar dilemma. If most of the data is reserved for training then we can have little confidence in the statistical stability of the test. On the other hand if most of the data is reserved for testing then the classifier is poorly trained. There is no definitive solution although there are more options available to us than just partitioning the data once, typically using a 50-50 split.

In this experiment we chose to use a *leaving one out* test/training strategy. With $n$ data items the data set is partitioned $n$ times. Each partition contains one data item in the test set and the remaining $n-1$ data items in the training set. The experiment is then repeated once on each partition, each partition containing a different data item in the test set. Although time consuming this approach has the advantage of including each data item in both the test and training sets.

Results

The recognition results for the three classifiers are presented in table 1. As would be expected the performance of the nn classifier based on the lpc data is significantly better than that based on the STFT power spectrum. Most interesting however is the performance of the PW covariance classifier which outperformed both the STFT and lpc based classifiers. This was despite the fact that this classifier used only one feature vector per transition, albeit a large one, while the other two techniques required multiple vectors to characterise each transition.

Table 1 Recognition performance of the 3 nn classifiers

| Recogniser | Recognition Performance |
|---|---|
| PW covariance | 91% |
| LPC | 87.5% |
| STFT | 82.5% |

## SUMMARY

Representing the speech signal by its power spectrum does not necessarily involve a loss of information about the signal even though the phase spectrum has been thrown away. If the power spectrum has been sampled at a sufficient rate in both time and frequency then signal recovery to within a constant phase factor is trivial[Altes 1980]. In current speech recognition systems however the power spectrum is typically undersampled in time by a factor of 100 or more. Consequently in the general case the orginal signal is no longer recoverable and information has been lost. The question we have attempted to provide an answer to in this paper is, is this information loss significant from the point of view of speech recognition? To answer this we have to be able to identify exactly what information has been thrown away by undersampling the power spectrum. This has been done by reformulating the problem in the pseudo-Wigner domain. Here we can show that an undersampled power spectrum is equivalent to representing the PWD in terms of its weighted first order statistics. The information thrown away then is contained in the higher order statistics of the PWD. The fact that we have demonstrated improved consonant recognition based on the covariance of the PWD suggests that a more accurate representation of the original speech signal may well produce better recognition results. In other words current speech recognition systems based on a short time spectral representation may well be discarding useful discriminant information contained in the input speech.

## REFERENCES

Allard. J.F., Valiere, J.C. and Bourdier, R. (1988) *Broadband Signal Analysis With The Smoothed Pseudo-Wigner Distribution*, J.Acoust.Soc.Am. 83(3), 1041-1044.

Altes, R.A. (1980) *Detection, Estimation, and Classification With Spectrograms*, J.Acoust. Soc.Am. 67(4), 1232-1246.

Boudreaux-Bartels, (1984) *Time Frequency Signal Processing Algorithms: Analysis and Synthesis Using Wigner Distributions*, PHD Thesis, Rice University

Claasen, T.A.C.M and Mecklenbraüker, W.F.G. (1980) *The Wigner Distribution - a Tool For Time Frequency Signal Analysis*, Philips J. Research, 35, Part 1:217-250, Part 2:276-300, Part 3:373-389.

Claasen, T.A.C.M and Mecklenbraüker, W.F.G. (1983) *The Aliasing Problem In Discrete-Time Wigner Distributions*, IEEE Trans. ASSP, Oct., 1067-1072

Hlawatsch, F. (1984) *Interference Terms In The Wigner Distribution*, Digital Signal Processing, eds. Cappellini, V. and Constantinides A.G. (North Holland, Amsterdam), 363-367.

Liberman, A.M. et al. (1956) *Journal of Experimental Psychology*, vol.52.

Lowe, D., Tomlinson, M.J. and Moore, R.K. (1986) *The Wigner Distribution as a Speech Processing Tool*, Proc.Inst.Acoust. Autumn Conf. Windermere, 97-102.

Priestley, M.B. (1981) *Spectral Analysis and Time Series Analysis*, (Academic Press).

Stevens, K.N and Blumstein, S.E. (1978) *Invariant Cues For Place of Articulation in Stop Consonants*, J.Acoust.Soc.Am. 64, 1358-1386.

Silverman, H.F. and Lee Yi-Teh (1987) *On The Spectrographic Representation of Rapidly Time-Varying Speech*, Computer Speech and Language, 2, 63-86.