TR-I-0164

Time-Frequency Spectral Estimation of Speech - The Role of Variance In Estimator Performance

D. Rainton

Abstract

In recent years there has been a growing interest amongst the speech research community into the use of spectral estimators which circumvent the traditional quasi-stationary assumption and provide greater time-frequency(t-f) resolution than conventional spectral estimators, such as the short time Fourier power spectrum (STFPS). One distribution in particular, the Wigner distribution (WD), has attracted considerable interest. However, experimental studies have indicated that, despite its improved t-f resolution, employing the WD as the front end of a speech recognition system actually reduces recognition performance; only by explicitly re-introducing t-f smoothing into the WD are recognition rates improved. This paper, by making explicit the role of the variance of the spectral estimator, provides a theoretical explanation for these previous experimental findings.

1 Introduction

Most researchers would agree that the speech signal is best represented for recognition purposes as a joint function of both time and frequency, or a parameterisation thereof.

Such a joint representation is both intuitively and biologically plausible. Biological studies for instance, have revealed that the ear acts essentially as a type of spectral analyser [10]. Thus by implication the spectral shape of the speech signal must contain important cues as to its information content. However, a spectral representation alone results in poor recognition performance, despite the fact that it may be a complete description of the signal (complete in the sense that the original acoustic waveform is uniquely recoverable). It is clear that the temporal ordering of speech events is also important for understanding the orthographic content of the acoustic waveform. Thus any useful representation must have both time and frequency dimensions. However, despite its obvious intuitive appeal, the mathematical description of temporal frequency variation has proven surprisingly difficult. This difficulty has arisen from the fact that the properties we would ideally require of such distributions [6] are mutually inconsistent. As a consequence of this fact, numerous t-f descriptions [5] have been proposed over the years; each satisfying only a non-conflicting subset of the ideal property set.

One of the first requirements of any t-f spectral estimation task is the selection of an appropriate estimator. In speech recognition this selection is largely based on previous empirical experience. One area of spectral analysis which has achieved impressive empirical results is that of perceptual modelling, *i.e.* modifying a given estimator to imitate the the way in which we believe that the human ear works. Yet, while such auditory modelling schemes have undoubtedly led to improved recognition performance, the question as to why this should be the case is left unanswered. Simply saying that one spectral estimator is better than another, because it models the behaviour of the human auditory system is, in our view, not enough. After all, a hidden Markov model, for example, makes no attempt to model the human brain, so why should it perform well with a human auditory style front end? Clearly there is something fundamentally right about these perceptually based spectral estimators, since they provide improved performance when coupled with almost any type of pattern recognition algorithm. If we could determine in an objective fashion exactly what they are optimising, then perhaps we could hope to improve upon them.

In recent years several authors [14] [15] have attempted to improve recognition performance by employing "high resolution" spectral estimators, such as the Wigner distribution (WD). Their argument being that conventional quasi-stationary estimators, such as the short time Fourier power spectrum (STFPS), introduce too much t-f "blurring" into the speech spectrum, thus masking important time varying features. However, in many cases, the use of "high resolution" estimators such as the WD has actually reduced recognition performance; only

by explicitly reintroducing t-f smoothing into the spectrum have recognition rates been increased. In this paper we argue that this is due to the higher variance of the "high resolution" estimators. To demonstrate the effect of variance on the recognition process we concentrate here on one particular family of t-f spectral estimators, namely the t-f smoothed WD's.

2 The T-F Smoothed Wigner Distribution

The mathematical properties of the WD have already been well documented elsewhere [1][2], consequently this section contains only a minimal introduction. The WD of the harmonisable stochastic process $S(t)$ is defined by the equation [8]

$$W_S(t,\omega) = \int_{-\infty}^{+\infty} R_S(t,\tau)e^{-j\omega\tau}d\tau \tag{1}$$

where $R_S(t,\tau)$ is the time-varying covariance kernel

$$R_S(t,\tau) = E[S(t+\tau/2)S^*(t-\tau/2)] \tag{2}$$

and $E[\cdot]$ denotes expectation over an ensemble of realisations. In practice ensemble averages are rarely available. For non-stationary random processes, such as speech, this necessitates replacing ensemble averages with local time averages. However, since speech is not a quasi correlation-ergodic process this assumption introduces a bias into the resulting spectral estimates. These biased spectral estimates are defined by the equation

$$\hat{W}_S(t,\omega;\phi) = \int_{-\infty}^{+\infty} \hat{R}_S(t,\tau;\psi)e^{-j\omega\tau}d\tau \tag{3}$$

where $\hat{R}_S(t,\tau;\psi)$ is in turn defined by the equation

$$\hat{R}_S(t,\tau;\psi) = \int_{-\infty}^{+\infty} \psi(\tau_1,\tau)S(t+\tau_1+\tau/2)S^*(t+\tau_1-\tau/2)d\tau_1 \tag{4}$$

and $\psi(\tau_1,\tau)$ is a window function, usually of finite spread and monotonic decreasing away from the origin. The window $\psi(\tau_1,\tau)$ is related to a corresponding t-f window $\phi(t,\omega)$ by a 1-D Fourier transform, *i.e.*

$$\phi(t,\omega) = \int_{-\infty}^{+\infty} \psi(t,\tau)e^{-j\omega\tau}d\tau. \tag{5}$$

For the purposes of this paper we assume that $\phi(t,\omega)$ is a non-orientated 2D gaussian window function, *i.e.*

$$\phi(t,\omega) = \frac{1}{\sigma_t\sigma_w}\exp(\frac{-t^2}{2{\sigma_t}^2})\exp(\frac{-\omega^2}{2{\sigma_w}^2}). \tag{6}$$

This window function is of particular interest, as for an appropriate choice of σ_t and σ_ω, the smoothed WD is identical to the STFPS [13]. Note that the normalising factor $1/\sigma_t\sigma_\omega$ is chosen such that $\phi(t,\omega)$ satisfies the equation

$$\int_{-\infty}^{\infty}\int \phi(t,\omega)d\omega dt = 2\pi. \tag{7}$$

This particular normalisation being chosen to ensure the equality

$$\int_{-\infty}^{\infty}\int E[\hat{W}_S(t,\omega;\phi)]dtd\omega = \int_{-\infty}^{\infty}\int W_S(t,\omega)dtd\omega. \tag{8}$$

From Moyal's formula [1] it can be seen that equation 8 implies that the expected energy of the smoothed spectral estimates $\hat{W}_S(t,\omega;\phi)$ is equal to that of $W_S(t,\omega)$.

Two important properties of the window function $\phi(t,\omega)$ are its time spread $(\Delta T)^2$ and its frequency spread $(\Delta B)^2$, defined respectively by equations 9 and 10, *i.e.*

$$(\Delta T)^2 = \frac{1}{E_\phi}\int_{-\infty}^{\infty}\int t^2|\phi(t,\omega)|^2 d\omega dt \tag{9}$$

$$(\Delta B)^2 = \frac{1}{2\pi E_\phi}\int_{-\infty}^{\infty}\int \omega^2|\phi(t,\omega)|^2 dtd\omega \tag{10}$$

where E_ϕ is defined as

$$E_\phi = \int_{-\infty}^{\infty}|\phi(t,\omega)|^2 dtd\omega. \tag{11}$$

In the case of the non-orientated 2D gaussian defined in equation 6, the time and frequency spreads of the smoothing window are directly proportional to σ_t and σ_ω respectively. Thus σ_t and σ_ω can be regarded as measures of the locality of the window function in time and frequency respectively.

3 Variance of $\hat{W}_S(t,\omega;\phi)$ as a Function of the T-F Spread of the Smoothing Window $\phi(t,\omega)$

One important, though often ignored, performance measure that can be attributed to the t-f smoothed estimates $\hat{W}_S(t,\omega;\phi)$, is their variance, defined as

$$\text{var}[\hat{W}_S(t,\omega;\phi)] = E[(\hat{W}_S(t,\omega;\phi) - E[\hat{W}_S(t,\omega;\phi)])^2]. \tag{12}$$

In order to make an analysis of equation 12 mathematically tractable we make the simplifying assumption that the speech signal $S(t)$ is a white zero mean complex analytic gaussian process. Obviously in practice this assumption is not satisfied by real speech. However, as long as the speech spectrum is relatively flat and smooth, then the analysis will still hold in an approximate sense. Since we are not attempting to draw any detailed conclusions from the analysis, but instead just looking for overall trends, we feel that the simplifying assumption is probably reasonable for this purpose. Given the above assumption then, the variance of the spectral estimates can be approximated by the equation

$$\text{var}\left[\hat{W}_S(t,\omega;\phi)\right] \approx \frac{1}{2\pi} S^2{}_t(\omega) \int\int_{-\infty}^{\infty} |\phi(t',\omega')|^2 \, dt' d\omega' \tag{13}$$

where $S_t(\omega)$ is the spectral density of the tangential stationary process approximating $S(t)$ at t, *i.e.*

$$S_t(\omega) = \int_{-\infty}^{+\infty} E[S(t+\tau/2)S(t-\tau/2)]e^{-j\omega\tau} d\tau. \tag{14}$$

Equation 13 is a straightforward generalisation of the discrete time case treated by Martin and Flandrin [8]. For the sake of completeness a detailed derivation is given in appendix A.

It is important to note that equation 13 implies that var $\left[\hat{W}_S(t,\omega;\phi)\right]$ is a monotonic decreasing function of the smoothing window parameters σ_t and σ_ω, since evaluating the integral on the right hand side of equation 13 gives

$$\begin{aligned}
\text{var}\left[\hat{W}_S(t,\omega;\phi)\right] &\approx \frac{1}{2\pi} S^2{}_t(\omega) \int\int_{-\infty}^{\infty} |\phi(t',\omega')|^2 \, dt' d\omega' \\
\text{var}\left[\hat{W}_S(t,\omega;\phi)\right] &\approx \frac{1}{2\pi} S^2{}_t(\omega) \frac{1}{(\sigma_t\sigma_\omega)^2} \int\int_{-\infty}^{\infty} \exp\left(\frac{-t'^2}{\sigma_t{}^2}\right) \exp\left(\frac{-\omega'^2}{\sigma_\omega{}^2}\right) dt' d\omega' \\
\text{var}\left[\hat{W}_S(t,\omega;\phi)\right] &\approx S^2{}_t(\omega) \frac{1}{2\sigma_t\sigma_\omega}.
\end{aligned} \tag{15}$$

In the next section we discuss the conflict between minimising estimator variance while simultaneously maximising the expected distance between spectral estimates corresponding to different speech sounds.

4 The Problem of Minimising the Inter-Sound Variance of Spectral Estimates While Simultaneously Maximising their Expected Intra-Sound Distance

We assume here a simplified model of the word generation process, where each orthographic word is modelled by a single unique stochastic process, which produces all acoustic realisations of that word.

Given two different stochastic processes, $\mathbf{U}(t)$ and $\mathbf{V}(t)$, corresponding to two different words in the recogniser vocabulary, we wish to choose $\phi(t,\omega)$ so as to minimise the probability of recognising estimates $\hat{W}_{\mathbf{V}}(t,\omega;\phi)$ as having been generated by the process $\mathbf{U}(t)$, and visa versa. In other words, we wish to choose $\phi(t,\omega)$ so as to maximise the word recognition rate. A closed form solution of the recognition rate as a function of $\phi(t,\omega)$ is clearly beyond our grasp. Consequently the following arguments are necessarily intuitive rather than rigorous.

Viewing the problem from a geometric point of view, each t-f spectral estimate will occupy a single point in an infinite dimensional Euclidean space. We would reasonably expect the spectral estimates associated with different words to produce different, identifiable, perhaps overlapping clusters in this space. An obvious optimisation criterion would be to choose $\phi(t,\omega)$ so as to maximise the spacing between the cluster centroids, while simultaneously minimising their spread.

Assuming a Euclidean distance measure, then the spacing between cluster centroids is by definition

$$\mathcal{D}(E[\hat{W}_{\mathbf{U}}], E[\hat{W}_{\mathbf{V}}]) = \int_{-\infty}^{\infty}\int |E[\hat{W}_{\mathbf{U}}(t,\omega;\phi)] - E[\hat{W}_{\mathbf{V}}(t,\omega;\phi)]|^2 dt d\omega. \quad (16)$$

Equation 16 can be re-expressed in the form (Appendix B)

$$\mathcal{D}(E[\hat{W}_{\mathbf{U}}], E[\hat{W}_{\mathbf{V}}]) = \int_{-\infty}^{\infty}\int |\Phi(\epsilon,\tau)|^2 |A_{\mathbf{U}}(\epsilon,\tau) - A_{\mathbf{V}}(\epsilon,\tau)|^2 d\epsilon d\tau \quad (17)$$

where $A_{\mathbf{U}}(\epsilon,\tau)$ is the Ambiguity Function of $\mathbf{U}(t)$ (equation 49), and $\Phi(\epsilon,\tau)$ is obtained from $\phi(t,\omega)$ via a 2D Fourier transform (equation 52). A detailed description of the Ambiguity function, widely used in radar signal processing, is given in [12]. Importantly appendix B shows that if $\phi(t,\omega)$ is an appropriately normalised (*i.e.* satisfies equation 7), real separable 2D gaussian, then $|\Phi(\epsilon,\tau)|^2$ must satisfy the inequality

$$|\Phi(\epsilon,\tau)|^2 \leq 1 \qquad \forall \epsilon,\tau. \tag{18}$$

Furthermore, $|\Phi(\epsilon,\tau)|^2$ is a monotonic everywhere (apart from $\epsilon = \tau = 0$) decreasing function of the t-f spread of $\phi(t,\omega)$. Thus $\mathcal{D}(E[\hat{W}_U], E[\hat{W}_V])$ must also be a monotonic *decreasing* function of the t-f spread of the smoothing window $\phi(t,\omega)$. Hence to increase the distance between cluster centroids we must reduce the spread of the smoothing window. However, this conflicts with the requirement of minimising the variance of the different individual clusters, since equation 13 shows that $\mathrm{var}[\hat{W}_S(t,\omega;\phi)]$ is a *decreasing* function of the smoothing window spread. Hence the best that we can hope for is to select a compromise value of window spread which minimises some combination of cluster centroid separation and variance. For example, one mathematically convenient optimisation criterion is to minimise the expected sum of the variance and squared bias, which is equivalent to minimising the mean squared error of the estimates [11]. For an adaptive approach to this mean squared error minimisation problem see for example [4] [11]. The mean squared error optimisation criterion of course assumes that we have no *a priori* information about the class to which a particular speech spectrum belongs. If such information is available then it makes more sense to smooth the spectra so as to maximise some discriminant function, rather than minimise residual error. However, such a discussion is beyond the scope of this paper, which seeks only to make explicit the role of estimator variance in the speech recognition process, thus clarifying the reasons for the failure of "high resolution" estimators, such as the WD.

5 Summary

The aim of this paper has been to explain, from a theoretical point of view, the reason why the use of "high resolution" estimators, such as the WD, has not led to improved speech recognition performance. We have argued that the reason for the poor performance of the WD is the relatively high variance of the resulting spectral estimates. Thus although quasi-stationary estimators, as as the STFPS, produce "blurring" of the spectral features, this is more than compensated for by the reduced variance of these resulting "blurred" spectra. Hence, while the cross terms of the WD [11] are a major disadvantage from the point of view of visual interpretation, the high estimator variance is the main problem from an automatic speech recognition point of view.

6 Appendix A - The Covariance of the T-F Smoothed Wigner Distribution - A Derivation

This appendix derives the following approximate relationship for the variance of the t-f smoothed WD $\hat{W}_{\mathrm{S}}(t,\omega;\phi)$, *i.e.*

$$\mathrm{var}\left[\hat{W}_{\mathrm{S}}(t,\omega;\phi)\right] \approx \frac{1}{2\pi}|S_{t_o}(\omega)|^2 \int\int_{-\infty}^{\infty} |\phi(t',\omega')|^2 \, d\omega' \, dt'. \tag{19}$$

when $\mathrm{S}(t)$ is a white zero mean *complex* analytic gaussian process.

The proof of the above equation is a straightforward extension of the discrete time proof given by Martin and Flandrin [7]. The derivation of equation 19 is as follows; for notational simplicity the shorthand notation $\hat{W}_{\mathrm{S}}(t_1,\omega_1;\phi) = W_1$ and $\hat{W}_{\mathrm{S}}(t_2,\omega_1;\phi) = W_2$ is used. Beginning with the expression for $\mathrm{cov}\,[W_1, W_2]$ gives

$$\mathrm{cov}\,[W_1, W_2] = E\,[(W_1 - E\,[W_1])(W_2 - E\,[W_2])^*]\,. \tag{20}$$

Substituting for W_1 and W_2 in eqn. 20 using the definition given in eqn. 3, *i.e.*

$$W_n \stackrel{def}{=} \int\int_{-\infty}^{\infty} \psi(\tau_1,\tau)\mathrm{S}(t+\tau_1+\tau/2)\mathrm{S}^*(t+\tau_1-\tau/2)e^{-j\omega\tau}\, d\tau\, d\tau_1 \tag{21}$$

gives

$$\mathrm{cov}\,[W_1, W_2] = \int\int\int\int_{-\infty}^{\infty} \psi(\tau_3,\tau_1)\psi^*(\tau_4,\tau_2)e^{-j(\omega_1\tau_1-\omega_2\tau_2)}\mathrm{cov}\,[st^*, uv^*]\, d\tau_1\, d\tau_2\, d\tau_3\, d\tau_4 \tag{22}$$

where

$$\begin{aligned} s &= \mathrm{S}(t_1+\tau_3+\tau_1/2) \qquad t = \mathrm{S}(t_1+\tau_3-\tau_1/2) \\ u &= \mathrm{S}(t_2+\tau_4+\tau_2/2) \qquad v = \mathrm{S}(t_2+\tau_4-\tau_2/2) \end{aligned} \tag{23}$$

Using the relationship (originally formulated by Isserlis [3])

$$\mathrm{cov}\,[st^*, uv^*] = E\,[st^*u^*v] - E\,[st^*]\,E\,[u^*v] \tag{24}$$

then if s,t,u,v are *real* joint Gaussian variates with zero averages

$$\operatorname{cov}[st^*, uv^*] = E[su^*]\,E[t^*v] + E[sv]\,E[t^*u^*]. \tag{25}$$

This result is easily derived from the characteristic function [9]. If however s, t, u, v are complex analytic zero mean gaussian variates it is straightforward to show that eqn. 25 reduces to

$$\operatorname{cov}[st^*, uv^*] = E[su^*]\,E[t^*v]. \tag{26}$$

This is because $E[wx] = 0$ where wx are zero mean analytic gaussian variables [9]. Thus for complex analytic zero mean gaussian variables $\operatorname{cov}[W_1, W_2]$ is simply

$$\operatorname{cov}[W_1, W_2] = A \tag{27}$$

where

$$A = \int_{-\infty}^{\infty}\int\int\int \psi(\tau_3, \tau_1)\psi^*(\tau_4, \tau_2)e^{-j(\omega_1\tau_1 - \omega_2\tau_2)}E[su^*]\,E[t^*v]\,d\tau_1\,d\tau_2\,d\tau_3\,d\tau_4 \tag{28}$$

To evaluate A expand $E[su^*]\,E[t^*v]$

$$\begin{aligned} E[su^*]\,E[t^*v] \;=\; & E[\mathrm{S}(t_1 + \tau_3 + \tau_1/2)\mathrm{S}^*(t_2 + \tau_4 + \tau_2/2)] \\ & E[\mathrm{S}^*(t_1 + \tau_3 - \tau_1/2)\mathrm{S}(t_2 + \tau_4 - \tau_2/2)] \end{aligned} \tag{29}$$

and temporarily substitute for $\tau_1, \tau_2, \tau_3, \tau_4$ using

$$\begin{aligned} \tau' &= a + b & \tau'' &= a - b \\ t' &= (c + d)/2 & t'' &= (c - d)/2 \end{aligned} \tag{30}$$

where

$$\begin{aligned} a &= t_1 - t_2 + \tau_3 - \tau_4 & b &= (\tau_1 - \tau_2)/2 \\ c &= t_1 + t_2 + \tau_3 + \tau_4 & d &= (\tau_1 + \tau_2)/2. \end{aligned} \tag{31}$$

Equation 29 can now be written in the form

$$E[su^*]\,E[t^*v] = E[\mathrm{S}(t' + \tau'/2)\mathrm{S}^*(t' - \tau'/2)]\,E^*[\mathrm{S}(t'' + \tau''/2)\mathrm{S}^*(t'' - \tau''/2)]. \tag{32}$$

Making the quasi-stationary assumption

$$E\left[\mathrm{S}(t_o+\tau/2)\mathrm{S}^*(t_o-\tau/2)\right] \approx E\left[\mathrm{S}(t'+\tau/2)\mathrm{S}^*(t'-\tau/2)\right] \approx E\left[\mathrm{S}(t''+\tau/2)\mathrm{S}^*(t''-\tau/2)\right] \quad (33)$$

where $t_o = (t_1+t_2)/2$ gives

$$E\left[su^*\right]E\left[t^*v\right] \approx E\left[\mathrm{S}(t_o+\tau'/2)\mathrm{S}^*(t_o-\tau'/2)\right] E^*\left[\mathrm{S}(t_o+\tau''/2)\mathrm{S}^*(t_o-\tau''/2)\right]. \quad (34)$$

Expressing the quasi-stationary covariance $E\left[(\mathrm{S}(t_o+\tau/2)\mathrm{S}^*(t_o-\tau/2)\right]$ in terms of the Fourier transform of the stationary process $\mathrm{S}_{t_o}(t)$, tangential to $\mathrm{S}(t)$ in t_o, *i.e.*

$$E\left[(\mathrm{S}(t_o+\tau/2)\mathrm{S}^*(t_o-\tau/2)\right] \approx \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{t_o}(\Omega)e^{j\Omega\tau}\,d\Omega \quad (35)$$

gives

$$E\left[su^*\right]E\left[t^*v\right] = \frac{1}{4\pi^2}\int\!\!\int_{-\infty}^{\infty} S_{t_o}(\Omega_1)S_{t_o}(\Omega_2)e^{j(\tau'\Omega_1-\tau''\Omega_2)}\,d\Omega_1\,d\Omega_2. \quad (36)$$

Substituting for $E\left[su^*\right]E\left[t^*v\right]$ in eqn. 28 using eqn. 36, rearranging factors and expanding τ', τ'' using equations 31 and 31

$$A \approx \frac{1}{4\pi^2}\int\!\!\int_{-\infty}^{\infty} \left\{\int\!\!\int_{-\infty}^{\infty} \psi(\tau_3,\tau_1)e^{-j\tau_3(\Omega_2-\Omega_1)}e^{-j\tau_1(\omega_1-[\Omega_1+\Omega_2]/2)}\,d\tau_1\,d\tau_3\right\} \left\{\int\!\!\int_{-\infty}^{\infty} \psi(\tau_4,\tau_2)e^{-j\tau_4(\Omega_2-\Omega_1)}e^{-j\tau_2(\omega_2-[\Omega_1+\Omega_2]/2)}\,d\tau_2\,d\tau_4\right\}^* S_{t_o}(\Omega_1)S_{t_o}(\Omega_2)e^{-j(\Omega_1-\Omega_2)(t_1-t_2)}\,d\Omega_1\,d\Omega_2 \quad (37)$$

Defining $\Phi(\Omega_a,\Omega_b)$ to be 2D Fourier transform of $\psi(\tau_a,\tau_b)$,*i.e.*

$$\Phi(\Omega_a,\Omega_b) \stackrel{def}{=} \int\!\!\int_{-\infty}^{\infty} \psi(\tau_a,\tau_b)e^{-j\Omega_a\tau_a}e^{-j\Omega_b\tau_b}d\tau_a d\tau_b \quad (38)$$

then substituting for the braced integrals in eqn. 37 using eqn. 38 and performing the co-ordinate transformation:- $\Omega' = (\Omega_2-\Omega_1)/2, \Omega'' = (\Omega_1+\Omega_2)/2$. The modulus of the Jacobian is 2 and hence $d\Omega_1\,d\Omega_2 \rightarrow 2d\Omega'\,d\Omega''$ giving

$$A \approx \frac{1}{2\pi^2} \iint_{-\infty}^{\infty} \left\{\Phi(2\Omega', \omega_1 - \Omega'') e^{-j2\Omega' t_1}\right\} \left\{\Phi^*(2\Omega', \omega_2 - \Omega'') e^{+j2\Omega' t_2}\right\} S_{t_o}(\Omega'' - \Omega') S^*_{t_o}(\Omega'' + \Omega')\, d\Omega'\, d\Omega''. \quad (39)$$

Defining $\Phi(\Omega, \omega)$ as a 1D Fourier transform of $\phi(t, \omega)$, *i.e.*

$$\Phi(\Omega, \omega) \stackrel{def}{=} \int_{-\infty}^{\infty} \phi(t, \omega) e^{-j\Omega t} dt \quad (40)$$

then eqn. 39 can be re-written as

$$A \approx \frac{1}{2\pi^2} \int\int_{-\infty}^{\infty}\int\int S_{t_o}(\Omega'' - \Omega') S^*_{t_o}(\Omega'' + \Omega') e^{-j2\Omega'(t' - t'')} \phi(t' - t_1, \omega_1 - \Omega'') \phi^*(t'' - t_2, \omega_2 - \Omega'')\, d\Omega'\, d\Omega''\, dt'\, dt''. \quad (41)$$

Since S(t) is a white noise process, $S_{t_o}(\omega)$ is flat and eqn. 41 is equivalent to

$$A \approx \frac{1}{2\pi^2} \int\int_{-\infty}^{\infty}\int\int |S_{t_o}(\Omega'')|^2 \left\{\int_{-\infty}^{\infty} e^{-j2\Omega'(t' - t'')}\, d\Omega'\right\} \phi(t' - t_1, \omega_1 - \Omega'') \phi^*(t'' - t_2, \omega_2 - \Omega'')\, d\Omega''\, dt'\, dt''. \quad (42)$$

Noting that $\int_{-\infty}^{\infty} e^{-j2\Omega'(t' - t'')}\, d\Omega'$ is equal to $\pi\delta(t' - t'')$ then eqn. 42 reduces to

$$A \approx \frac{1}{2\pi} \int\int_{-\infty}^{\infty} |S_{t_o}(\Omega'')|^2 \phi(t' - t_1, \omega_1 - \Omega'') \phi^*(t' - t_2, \omega_2 - \Omega'')\, d\Omega''\, dt'. \quad (43)$$

Using exactly the same set of arguments it can be shown that

$$B \approx \frac{1}{2\pi} \int\int_{-\infty}^{\infty} |S_{t_o}(\Omega'')|^2 \phi(t' - t_1, \omega_1 - \Omega'') \phi^*(t' - t_2, \omega_2 + \Omega'')\, d\Omega''\, dt'. \quad (44)$$

Hence, when the random process S(t) is a white zero mean *complex* analytic gaussian process, the covariance of its smoothed WD is approximately equal to

$$\text{cov}\left[\hat{W}_{\mathbf{S}}(t_1,\omega_1;\phi)\hat{W}_{\mathbf{S}}(t_2,\omega_2;\phi)\right] \approx$$
$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\int |S_{t_o}(\Omega'')|^2\phi(t'-t_1,\omega_1-\Omega'')\phi^*(t'-t_2,\omega_2-\Omega'')\,d\Omega''\,dt'. \quad (45)$$

Letting t_1 and t_2 equal t, and ω_1 and ω_2 equal ω, we obtain the following approximate relationship for the variance of the WD, *i.e.*

$$\text{var}\left[\hat{W}_{\mathbf{S}}(t,\omega;\phi)\right] \approx$$
$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\int |S_t(\Omega'')|^2\phi(t'-t,\omega-\Omega'')\phi^*(t'-t,\omega-\Omega'')\,d\Omega''\,dt'. \quad (46)$$

Finally, using the fact the $S_{t_o}(\omega)$ is flat in frequency, then we obtain the desired result

$$\text{var}\left[\hat{W}_{\mathbf{S}}(t,\omega;\phi)\right] \approx \frac{1}{2\pi}|S_t(\omega|^2\int_{-\infty}^{\infty}\int |\phi(t',\omega')|^2\,d\omega'\,dt'. \quad (47)$$

7 Appendix B - Changes in $\mathcal{D}(E[\hat{W}_{\mathbf{U}}], E[\hat{W}_{\mathbf{V}}])$ as a function of the t-f spread of $\phi(t,\omega)$

This appendix shows that if $\phi(t,\omega)$ is a real separable gaussian window, which satisfies the normalisation criterion of equation 7, then $\mathcal{D}(E[\hat{W}_{\mathbf{U}}], E[\hat{W}_{\mathbf{V}}])$ must be a *monotonic decreasing* function of the t-f spread of $\phi(t,\omega)$, irrespective of the choice $\mathbf{U}(t)$ and $\mathbf{V}(t)$.

Beginning with the definition of $\mathcal{D}(E[\hat{W}_{\mathbf{U}}], E[\hat{W}_{\mathbf{V}}])$ given in equation 16, *i.e.*

$$\mathcal{D}(E[\hat{W}_{\mathbf{U}}], E[\hat{W}_{\mathbf{V}}]) = \int\int_{-\infty}^{\infty} |E[\hat{W}_{\mathbf{U}}(t,\omega;\phi)] - E[\hat{W}_{\mathbf{V}}(t,\omega;\phi)]|^2 dt d\omega \tag{48}$$

we re-formulate this equation in the Ambiguity plane. The Ambiguity function (AF), widely used in radar signal processing [12], is defined by the equation

$$A_{\mathbf{S}}(\epsilon,\tau) = \mathcal{F}(W_{\mathbf{S}}(t,\omega)) \tag{49}$$

where

$$\mathcal{F}(W_{\mathbf{S}}(t,\omega)) = \frac{1}{2\pi}\int\int_{-\infty}^{\infty} W_{\mathbf{S}}(t,\omega) e^{-j(\epsilon t - \omega\tau)} dt d\omega. \tag{50}$$

Similarly

$$\hat{A}(\epsilon,\tau;\Phi) = \mathcal{F}(\hat{W}(t,\omega;\phi)) \tag{51}$$

where

$$\Phi(\epsilon,\tau) = \mathcal{F}(\phi(t,\omega)). \tag{52}$$

Making use of Parseval's relationship, and the equality

$$A_{\mathbf{S}}(\epsilon,\tau)\Phi(\epsilon,\tau) = \hat{A}_{\mathbf{S}}(\epsilon,\tau;\Phi) \tag{53}$$

equation 48 can be re-expressed in the form

$$\mathcal{D}(E[\hat{W}_{\mathbf{U}}], E[\hat{W}_{\mathbf{V}}]) = \int\int_{-\infty}^{\infty} |\Phi(\epsilon,\tau)|^2 |A_{\mathbf{U}}(\epsilon,\tau) - A_{\mathbf{V}}(\epsilon,\tau)|^2 d\epsilon d\tau. \tag{54}$$

If $\phi(t,\omega)$ is a real separable 2D gaussian smoothing window, defined by the equation

$$\phi(t,\omega) = \frac{1}{\sigma_t \sigma_w} \exp(\frac{-t^2}{2{\sigma_t}^2}) \exp(\frac{-\omega^2}{2{\sigma_w}^2}), \tag{55}$$

which can be shown to satisfy the required normalisation of equation 7, then it is straightforward to derive that

$$\Phi(\epsilon,\tau) = \exp(\frac{-\epsilon^2{\sigma_t}^2}{2})\exp(\frac{-\tau^2{\sigma_w}^2}{2}). \tag{56}$$

Thus $\Phi(\epsilon,\tau)$ will also be a real separable 2D gaussian which satisfys the inequalities

$$\Phi(0,0) = 1 \qquad \text{and} \qquad \Phi(\epsilon,\tau) \leq 1 \quad \forall\epsilon,\tau. \tag{57}$$

Furthermore, since $|\Phi(\epsilon,\tau)|^2$ must also be less or equal to 1 for all ϵ,τ, then equation 54 implies the following inequality, *i.e.*

$$\mathcal{D}(E[\hat{W}_{\mathbf{U}}], E[\hat{W}_{\mathbf{V}}]) \leq \mathcal{D}(W_{\mathbf{U}}, W_{\mathbf{V}}). \tag{58}$$

Or in other words, the distance between the expected values of the smoothed spectral estimates must be less than the distance between the unsmoothed estimates. Furthermore, $|\Phi(\epsilon,\tau)|^2$ is a monotonic decreasing function of σ_t and σ_w, for all ϵ,τ. Hence $\mathcal{D}(E[\hat{W}_{\mathbf{U}}], E[\hat{W}_{\mathbf{V}}])$ must be monotonic decreasing function of window spread, and our original claim is proven.

References

[1] T. A. S. M. Classen and W. F. G. Mecklenbräuker. The Wigner Distribution - A Tool for Time-Frequency Analysis Part 1: Continuous-Time Signals. *Philips J. Res.*, pages 217–250, 1980.

[2] T. A. S. M. Classen and W. F. G. Mecklenbräuker. The Wigner Distribution - A Tool for Time-Frequency Analysis Part 2: Discrete-Time Signals. *Philips J. Res.*, pages 276–300, 1980.

[3] L. Isserlis. On a Formula for the Product-Moment Coefficient of Any Order of a Normal Frequency Distribution in Any Number of Variables. *Biometrica*, 12:134–139, 1918.

[4] N. C. Sedgwick J. S. Bridle. A Method for Segmenting Acoustic Speech Patterns, With Applications to Automatic Speech Recognition. *ICASSP*, 1977.

[5] D. Lowe. Joint Representations in Quantum Mechanics and Signal Processing Theory: Why a Probability Function of Time and Frequency is Disallowed. Technical Report 4017, Royal Signals and Radar Establishment, 1986.

[6] R. M. Loynes. On the Concept of the Spectrum for Non-Stationary Processes. *J. Roy. Statist. Soc. ser. B, vol.30*, pages 1–20, 1968.

[7] W. Martin and P. Flandrin. *Analysis of Non-Stationary Processes: Short Time Periodograms Versus a Pseudo Wigner Estimator*, pages 455–458. Elsevier Science Publishers, 1983.

[8] W. Martin and P. Flandrin. Wigner-Ville Spectral Analysis of Nonstationary Processes. *IEEE Trans. Acoust., Speech, Signal Processing*, pages 1461–1470, 1985.

[9] A. Papoulis. *The Fourier Integral and Its Applications.* McGraw-Hill, 1962.

[10] J. O. Pickles. *An Introduction to the Physiology of Hearing.* Academic Press, 1982.

[11] D. Rainton. Time-Frequency Spectral Estimation of Speech. Technical Report CUED/F-INFENG/TR.39, Cambridge University Engineering Department, 1990.

[12] A. W. Rihaczek. *Principles of High Resolution Radar.* McGraw-Hill Book Company, 1969.

[13] M. D. Riley. Beyond Quasi-Stationarity: Designing Time-Frequency Representations for Speech Signals. In *Proc ICASSP*, pages 657–660, 1987.

[14] E. F. Velez and R. G. Absher. Transient Analysis of Speech Signals using the Wigner Time-Frequency Representation. In *Proc ICASSP*, pages 2242–2245, 1989.

[15] J. Wilbur and F.J. Taylor. Consistent Speaker Identification via Wigner Smoothing Techniques. In *Proc ICASSP*, pages 591–594, 1988.