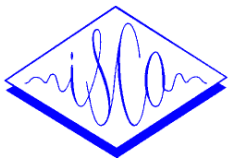

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Time-Frequency Spectral Analysis of Speech

David Rainton
ATR Interpreting Telephony Research Laboratories
S. J. Young
Cambridge University Department of Engineering

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 spectra (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. By re-formulating the spectral estimation problem in terms of a bias variance optimisation task, we provide an explanation for these previous experimental findings.

A practical adaptive smoothing algorithm is introduced, which attempts to match the degree of smoothing introduced into the WD with the time varying quasi-stationary regions within the speech waveform. The recognition performance of the resulting adaptively smoothed estimator is found to be comparable to that of conventional filterbank estimators, yet the average temporal sampling rate of the resulting spectral vectors is reduced by around a factor of ten.

INTRODUCTION

It is generally agreed 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. Thus by implication the spectral shape of the speech signal must contain important cues as to its information content. However, a spectral representation alone is insufficient, 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[1] form a mutually inconsistent set. As a consequence numerous t-f descriptions[2] have been proposed over the years; each satisfying only a non-conflicting subset of the ideal property set.

Historically, the most popular family of t-f spectral estimators have been the well known STFPS. Today however, it is well understood that this family of estimators is simply a subset of a more general class of estimators, known as the t-f smoothed Wigner distributions (TFSWD), this family itself belonging to a larger group of distributions known as Cohen's class[2]. This aim of this paper is the formulation of an optimal adaptive TFSWD spectral estimation strategy; adaptive because the degree of t-f smoothing introduced into the signal spectrum must continually adapt to match the changing local signal statistics. This is in sharp contrast to conventional spectral estimation techniques which typically apply a single fixed estimator to all the speech data, irrespective of the local signal characteristics.

T-F SMOOTHED SPECTRAL ESTIMATES

For the purposes of this paper the speech signal $\mathbf{S}(t)$ is treated as a harmonisable random stochastic process. The resulting TFSWD are then defined as follows[3]

$$\hat{W}_{\mathbf{S}}(t,\omega;\phi) = \int \hat{R}_{\mathbf{S}}(t,\tau;\psi)e^{-j\omega\tau}d\tau, \tag{1}$$

where $\hat{R}_{\mathbf{S}}(t,\tau;\psi)$ is equal to

$$\hat{R}_{\mathbf{S}}(t,\tau;\psi) = \int \psi(\tau_1,\tau)\mathbf{S}(t+\tau_1+\tau/2)\mathbf{S}^*(t+\tau_1-\tau/2)\,d\tau_1. \tag{2}$$

The windows ϕ and ψ are related by a 1D Fourier transform, *i.e.* $\phi(t,\omega) = \int \psi(t,\tau_1)e^{-j\omega\tau_1}d\tau_1$. Selection of a particular member of the family of TFSWD is determined by the choice of window function $\phi(t,\omega)$ (or equivalently $\psi(t,\omega)$).

It is important to fully appreciate that $\hat{W}_{\mathbf{S}}(t,\omega;\phi)$ are *estimates* [1] of the Wigner distribution (WD), the variance and bias of these estimates, for a given stochastic process, depending directly on the window function $\phi(t,\omega)$. The exact nature of this bias, variance relationship is examined in more detail later in the paper.

The WD itself is defined as follows[3]

$$W_{\mathbf{S}}(t,\omega) = \int R_{\mathbf{S}}(t,\tau)e^{-j\omega\tau}d\tau \tag{3}$$

where $R_{\mathbf{S}}(t,\tau)$ is the time-varying covariance kernel

$$R_{\mathbf{S}}(t,\tau) = E[\mathbf{S}(t+\tau/2)\mathbf{S}^*(t-\tau/2)] \tag{4}$$

and $E[\cdot]$ denotes expectation over an ensemble of realisations. It is straightforward to show that the WD and the expected value of the t-f smoothed estimates $\hat{W}_{\mathbf{S}}(t,\omega;\phi)$ are related via a t-f convolution[3], *i.e.*

$$E[\hat{W}_{\mathbf{S}}(t,\omega;\phi)] = \frac{1}{2\pi}\int\int W_{\mathbf{S}}(t_1,\omega_1)\phi(t-t_1,\omega-\omega_1)dt_1d\omega_1. \tag{5}$$

The rest of this paper is concerned with selection of an optimal t-f smoothing window function $\phi(t,\omega)$. To simplify the mathematics, we restrict our analysis to t-f smoothing window functions of the form

$$\phi(t,\omega) = \frac{1}{\rho_t\rho_\omega}\exp(\frac{-t^2}{2\rho_t{}^2})\exp(\frac{-\omega^2}{2\rho_\omega{}^2}) \tag{6}$$

[1] indicated by the hat notation

where the positive scalars ρ_t, ρ_ω are measures of the spread of the window in the time and frequency directions respectively. Thus the choice of window function is completely governed by selection of an appropriate set of values for these two parameters. Of particular interest is the case where ρ_t, ρ_ω are chosen to satisfy the equality $\rho_t\rho_\omega = 0.5$, since then the resulting spectrum can be shown to belong to the family of STFPS[4].

SOME GENERAL GUIDELINES FOR THE SELECTION OF $\phi(t,\omega)$

In recent years several authors have attempted to improve speech recognition performance by reducing the combined t-f spread of the smoothing window to below that required to produce a STFPS (*i.e.* $\rho_t\rho_\omega < 0.5$). The justification for this being that the t-f blurring introduced into the spectrum by the window function $\phi(t,\omega)$ may well obscure useful recognition features within the speech signal, particularly at consonant-vowel boundaries. Hence reducing the combined t-f window spread should improve recognition performance, since more spectral detail should result. However, the performance of these *high resolution* spectral estimators, as they are often referred to, has been unspectacular[5][6]. Indeed recognition rates have actually been reduced in many cases. This fact, coupled with the high computational burden associated with such estimators, has made them distinctly unattractive.

The reason for the reduced recognition performance associated with the reduced t-f window spread is explained only if we additionally take into account the variance of the resulting spectral estimates. Viewing the spectral estimation problem from a geometric point of view, each t-f spectral estimate will occupy a single point in an infinite dimensional Euclidean space. Given two different word stochastic processes, $\mathbf{U}(t)$ and $\mathbf{V}(t)$, each process producing all acoustic realisations of a corresponding orthographic form in the recogniser's lexicon, it would seem reasonable to expect the t-f spectral estimates associated with different words to produce different, identifiable, perhaps overlapping clusters in this Euclidean space. Clearly any selection of ρ_t and ρ_ω should be such as to maximise the spacing between the cluster centroids, $E[\hat{W}_\mathbf{U}]$ and $E[\hat{W}_\mathbf{V}]$, while simultaneously minimising their spreads, *i.e.* minimising $\mathrm{var}[\hat{W}_\mathbf{U}]$ and $\mathrm{var}[\hat{W}_\mathbf{V}]$.

THE SEPARATION BETWEEN EXPECTED SPECTRAL ESTIMATES AS A FUNCTION OF THE T-F SPREAD OF $\phi(t,\omega)$

In this section we claim that as the t-f spread of the smoothing window $\phi(t,\omega)$ is increased, so the expected Euclidian distance between spectral estimates generated from different word realisations must decrease.

A brief outline of the proof is as follows; defining $\mathcal{D}(E[\hat{W}_\mathbf{U}], E[\hat{W}_\mathbf{V}])$ as the Euclidean distance between the cluster centroids, $E[\hat{W}_\mathbf{U}]$ and $E[\hat{W}_\mathbf{V}]$, *i.e.*

$$\mathcal{D}(E[\hat{W}_\mathbf{U}], E[\hat{W}_\mathbf{V}]) = \int\int |E[\hat{W}_\mathbf{U}(t,\omega;\phi)] - E[\hat{W}_\mathbf{V}(t,\omega;\phi)]|^2 dt d\omega \tag{7}$$

then, making use of Parsival's relationship, it is possible to show that

$$\mathcal{D}(E[W_\mathbf{U}], E[W_\mathbf{V}]) = \int\int |\Phi(\epsilon,\tau)|^2 |A_\mathbf{U}(\epsilon,\tau) - A_\mathbf{V}(\epsilon,\tau)|^2 d\epsilon d\tau \tag{8}$$

where the Ambiguity function $A_\mathbf{S}(\epsilon,\tau)$ is related to the WD by a 2D Fourier transform, *i.e.*

$$A_\mathbf{S}(\epsilon,\tau) = \frac{1}{2\pi}\int\int W_\mathbf{S}(t,\omega) e^{-j(\epsilon t - \omega\tau)} dt d\omega. \tag{9}$$

A similar relationship exists between $\Phi(\epsilon,\tau)$ and $\phi(t,\omega)$, *i.e.*

$$\Phi(\epsilon,\tau) = \frac{1}{2\pi}\int\int \phi(t,\omega) e^{-j(\epsilon t - \omega\tau)} dt d\omega. \tag{10}$$

Assuming $\phi(t,\omega)$ to be a separable 2D gaussian of the form given in eqn. 6, which can be shown to satisfy the normalisation

$$\int\int \phi(t,\omega) dt d\omega = 2\pi, \tag{11}$$

then it is straightforward to show that $|\Phi(\epsilon,\tau)|^2$ is an everywhere (apart from at the origin where $\Phi(0,0) = 1$ forall ρ_t and ρ_ω) monotonic decreasing function of ρ_t and ρ_ω. Hence from eqn. 8, the Euclidian distance $\mathcal{D}(E[\hat{W}_\mathbf{U}], E[\hat{W}_\mathbf{V}])$ must also be monotonic decreasing function of ρ_t and ρ_ω. In other words, for any given pair of stochastic processes $\mathbf{U}(t)$ and $\mathbf{V}(t)$, the Euclidian distance between expected spectral estimates is a monotonic decreasing function of the t-f spread of the $\phi(t,\omega)$. A more detailed explanation of the above proof is given in[7].

VARIANCE OF THE SPECTRAL ESTIMATES AS A FUNCTION OF THE T-F SPREAD OF $\phi(t,\omega)$

The variance of the t-f smoothed estimates, defined as

$$\mathrm{var}[\hat{W}_\mathbf{S}(t,\omega;\phi)] = E[(\hat{W}_\mathbf{S}(t,\omega;\phi) - E[\hat{W}_\mathbf{S}(t,\omega;\phi)])^2] \tag{12}$$

can be show to satisfy the approximation

$$\mathrm{var}\left[\hat{W}_\mathbf{S}(t,\omega;\phi)\right] \approx \frac{1}{2\pi} {f_t}^2(\omega) \int\int |\phi(t',\omega')|^2 \, dt' d\omega' \tag{13}$$

where f_t is the spectral density of the tangential stationary process approximating $\mathbf{S}(t)$ at t. Equation 13 is a straightforward generalisation of the discrete time case treated by Martin and Flandrin[3]. A detailed derivation is given in[7]. It is straightforward to show from eqns 13 and 6 that the variance of the spectral estimates is a monotonic *decreasing* function of ρ_t and ρ_ω[7].

OPTIMAL SELECTION OF $\phi(t,\omega)$

Following the discussion in the earlier GENERAL GUIDELINES section we would ideally like to select the t-f spread of the smoothing window $\phi(t,\omega)$ so as to maximise the separation between the expected spectral estimates corresponding to different word stochastic processes, while at the same time minimising the variance of those estimates corresponding to the same word. However, in the previous two sections we have shown that both the variance and centroid separation are monotonic decreasing functions of the t-f window spread. Hence a conflict arises in simultaneously attempting to minimise one while maximising the other. The best that we can hope for is to select some compromise value of window spread which minimises some combination of cluster centroid separation and variance.

BIAS OF THE SPECTRAL ESTIMATES

One further measure of interest is the bias of the resulting spectral estimates, defined as

$$\mathrm{bias}[\hat{W}_\mathbf{S}(t,\omega;\phi)] = E[\hat{W}_\mathbf{S}(t,\omega;\phi)] - W_\mathbf{S}(t,\omega). \tag{14}$$

In particular it is possible to show, by a simple geometric argument, that the reduction in the average spectral distance due to t-f smoothing is constrained according to the equation,

$$\mathcal{D}(W_\mathbf{U}, W_\mathbf{V}) - \mathcal{D}(E[\hat{W}_\mathbf{U}], E[\hat{W}_\mathbf{V}]) \leq \int\int |\mathrm{bias}[\hat{W}_\mathbf{U}(t,\omega)]|^2 + |\mathrm{bias}[\hat{W}_\mathbf{V}(t,\omega)]|^2 dt d\omega. \tag{15}$$

By a mini-max argument, minimising the bias of the spectral estimates will maximise the minimum possible distance between their expected values.

The bias itself can be shown to satisfy the following approximation[7],

$$\mathrm{bias}\left[\hat{W}_\mathbf{S}(t,\omega;\phi)\right] \approx \frac{1}{4\pi}\mathcal{D}_t{}^2 W_\mathbf{S}(t,\omega) \int \left\{ \int {t_1}^2 \phi(t_1,\omega_1) dt_1 \right\} d\omega_1 + \frac{1}{4\pi}\mathcal{D}_\omega{}^2 W_\mathbf{S}(t,\omega) \int \left\{ \int {\omega_1}^2 \phi(t_1,\omega_1) d\omega_1 \right\} dt_1 \tag{16}$$

where $\mathcal{D}_x{}^r = \frac{\partial^r}{\partial x^r}$. Equation 16 implicitly assumes that $W_{\mathbf{S}}(t,\omega)$ is locally quadratic, *i.e.* is twice differentiable with a bounded second derivative, and that $\phi(t,\omega)$ is normalised according to eqn. 11.

Equation 16 shows that the bias (and so by implication the minimum possible distance between expected spectral estimates) depends not only on the t-f spread of the window $\phi(t,\omega)$, but also on $\mathcal{D}_t{}^2 W_{\mathbf{S}}(t,\omega)$ and $\mathcal{D}_\omega{}^2 W_{\mathbf{S}}(t,\omega)$, which are measures of the time and frequency curvatures of $W_{\mathbf{S}}(t,\omega)$ respectively. Hence the sharper the local curvature of the t-f spectrum, the greater the bias for a given choice of smoothing window.

Up to now we have implicitly assumed that $\phi(t,\omega)$ is t-f shift invariant, thus $W_{\mathbf{S}}(t,\omega)$ and $E[\hat{W}_{\mathbf{S}}(t,\omega)]$ are related via the t-f convolution in eqn. 5. However, eqns 15 and 16 suggest that a more optimal solution to the bias/variance tradeoff could be achieved if $\phi(t,\omega)$ were allowed to adapt its t-f spread to match the local curvature of the t-f spectrum. In other words the spread of the smoothing window should increase in the flatter regions and decrease in the more sharply curved regions of the t-f spectrum.

A PRACTICAL ADAPTIVE SMOOTHING ALGORITHM

Having outlined a general optimisation strategy in the first half of this paper we now pursue its practical implementation. The adaptive spectral estimation algorithm proposed here is a two stage process. First the spectral estimates $\hat{W}_{\mathbf{S}}(t,\omega;\phi)$ are computed using a t-f shift invariant window $\phi(t,\omega)$, whose time spread and frequency spread are chosen to remove glottal excitation effects from the resulting spectrum. Riley[4] has shown that an appropriate window is one whose time spread is equal to one pitch period and whose frequency spread is equal to the spacing between the spectral harmonics. This initial smoothing is based on the assumption that the glottal excitation signal contains no useful information about the orthographic content of an utterance, instead it simply adds to the variance of the spectral estimates.

The second stage of the algorithm involves adaptively smoothing $\hat{W}_{\mathbf{S}}(t,\omega;\phi)$ in time to produce $\hat{W}_{\mathbf{S}}(t,\omega;\phi\check{\phi})$, using a time *variant* gaussian smoothing window $\check{\phi}(t,t')$, *i.e.*

$$\hat{W}_{\mathbf{S}}(t,\omega;\phi\check{\phi}) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{W}_{\mathbf{S}}(t',\omega;\phi)\check{\phi}(t,t')\,dt' \tag{17}$$

where

$$\check{\phi}(t,t') = \frac{1}{\sigma(t)\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{t-t'}{\sigma(t)}\right)^2\right) \tag{18}$$

Note that $\check{\phi}(t,t')$ is not shift invariant since the variance of this smoothing window is a function of time.

The idea is to select $\sigma(t)$ so as to match the changing quasi-stationary periods within the speech signal, thus introducing maximum variance reduction, while introducing minimal additional bias. Thus for example, mid-way into a long vowel sound $\sigma(t)$ will be large, but at a consonant-vowel boundary it will be small. Conventional spectral estimators on the other hand make the blanket assumption that the speech signal comprises of fixed length quasi-stationary intervals. The assumed duration of these intervals being based on prior information about the expected signal statistics. In reality, it is more reasonable to expect $\sigma(t)$ to be some continuously changing function of time. However, to simplify the optimisation problem $\sigma(t)$ is constrained here to be piecewise linear, *i.e.*

$$\begin{aligned}
\sigma(t) &= \sigma_{i-1} + \tfrac{t-t_{i-1}}{t_i-t_{i-1}}(\sigma_i - \sigma_{i-1}) && t_{i-1} \le t \le t_i \quad \text{and} \quad t > t_1 \\
\sigma_i &= (t_i - t_{i-1})/\pi && 2 \le i \le N \\
\sigma(t) &= \sigma_1 && t \le t_1
\end{aligned} \tag{19}$$

where t_0 marks the start of the utterance and t_N marks its end (fig. 1).

In other words, $\sigma(t)$ is approximated by a set of N straight line segments. It is assumed that N is selected on an *a priori* basis. The σ_i values are defined as they are in eqn. 19 so that the resulting spectrum can be reasonably sampled at the set of time points t_i, while satisfying what can be roughly thought of as a generalised Nyquist criterion[7].

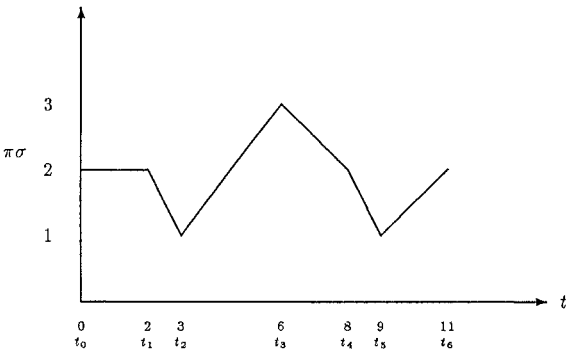

Figure 1: One possible form of $\sigma(t)$, satisfying the constraints given in eqn. 19, with N=6.

Thus instead of having to optimise an arbitrary function $\sigma(t)$, the problem is now reduced to one of finding an optimal set of $N-1$ breakpoints, *i.e.* the t_i values $1 \le i \le N-1$.

The optimal set of t_i values is determined to be that set which minimises the mean squared distance between $\hat{W}_{\mathbf{S}}(t,\omega;\phi\check{\phi})$ and $\hat{W}_{\mathbf{S}}(t,\omega;\phi)$. In practice this optimisation is accomplished by sampling the bark scaled t-f spectrum $\hat{W}_{\mathbf{S}}(t,\omega;\phi)$ at the Nyquist rate and then applying a dynamic programming algorithm to the resulting spectral vector sequence[7].

A brief outline of the algorithm is as follows; the vector sequence $\mathbf{s}_t$ is defined as the m-dimensional column vector obtained by sampling $\hat{W}_{\mathbf{S}}(t\Delta_t,\omega;\phi)$ at the frequency points $0, \Delta_\omega, \ldots, m\Delta_\omega$, *i.e.*

$$\mathbf{s}_t' = [s_{t,1}, \ldots, s_{t,i}, \ldots, s_{t,m}] \qquad \text{where} \quad s_{t,i} = \hat{W}_{\mathbf{S}}(t\Delta_t, (i-1)\Delta_\omega;\phi). \tag{20}$$

The sampling intervals, Δ_t and Δ_ω, are $\pi\sigma_t$ and $\pi\sigma_\omega$ respectively, so as to satisfy the Nyquist criterion[7]. $T\Delta_t$ seconds of speech produces the vector sequence $\mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_t, \ldots, \mathbf{s}_T$.

Let $F_n(t)$ be the minimum distortion introduced by smoothing the vector sequence $\mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_t$ using a discrete time equivalent of eqn. 19. For $n > 1$, the well known principle of optimality yields the recursive relationship

$$F_n(t) = \min_{n-1 \le t' < t} [F_{n-1}(t') + D(t',t)]. \tag{21}$$

The MSE distortion measure $D(t',t)$ is defined as

$$\begin{aligned}
D(t',t) &= \sum_{j=t'}^{t} ||\mathbf{s}_j - \bar{\mathbf{s}}_j||^2 \\
\bar{s}_{j,i} &= \sum_k \phi(j,k) s_{k,i} \\
\phi(j,k) &= \frac{1}{\sigma_j\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{(j-k)^2}{\sigma_j^2}\right) \\
\sigma_j &= \sigma_{t'} + \frac{j-t'}{t-t'}(\sigma_t - \sigma_{t'}) \\
\sigma_t &= \frac{t-t'}{\pi} \\
\text{if} \quad t' = 0 \quad \text{then} \quad \sigma_{t'} &= \sigma_t \quad \text{else} \quad \sigma_{t'} = \sigma_{t',N-1} \\
\sigma_{t,N} &= \sigma_t.
\end{aligned} \tag{22}$$

The starting point for the recursion is the computation of $F_1(t)$, which is defined as

$$F_1(t) = D(0,t). \tag{23}$$

This algorithm returns the optimal set of $N-1$ breakpoints, $t_1, \ldots, t_{N-1}$. The N optimally smoothed spectral vectors, $\mathbf{s}_{t_j} \quad 1 \le j \le N$, can now be recovered by computing

$$s_{t_j} = \sum_{k=0}^{T} \phi(t_j, k) s_k \qquad 1 \leq j \leq N \tag{24}$$

where $\phi(t_j, k)$ is defined in eqn. 22. N may be selected on an *a priori* basis, or may be chosen to ensure that the average MSE between corresponding vector sequences, obtained before and after adaptive smoothing, is below some pre-defined threshold.

EXPERIMENTAL DETAILS

The speech used for the following experiments comprised of the isolated digits zero to nine. The speech database contained one thousand four hundred isolated digits, obtained from nine female and twenty six male speakers. For each word realisation, the initial spectral analysis stage, employing a fixed t-f smoothing window, produced a sequence of 20 dimensional, bark scaled spectral vectors, one vector every 10ms. Half of the resulting vector sequences from each speaker were used for training the recogniser and the other half used for testing the recogniser.

Although the adaptive spectral vectors could in principle be used to train most types of conventional recogniser, including hidden Markov models (HMM) and neural networks (NN), a standard dynamic time warping (DTW) based recogniser was used for this experiment[8].

EXPERIMENTAL RESULTS

Three separate recognition experiments were conducted, all using an identical DTW pattern recognition algorithm[8], the only difference being the type of spectral vectors used. For a control experiment we used the spectral vectors produced prior to the second adaptive smoothing stage. The recognition rate obtained using this conventional style of vector was 97.7%. This figure should be compared against that obtained from experiments 1 and 2. Experiment 1 used the adaptively smoothed spectral vectors and experiment 2 used the adaptively smoothed vectors augmented with adaptive difference vectors (*i.e.* the i'th adaptive difference vector is defined as the difference between the i'th and $i+1$'th adaptively smoothed vectors). These latter two experiments were repeated using a range of values for N. The results of these two experiments are shown in table 1.

N	Exp1	Exp2
4	85.2%	88.4%
5	92.7%	92.7%
6	94.3%	94.2%
8	97.9%	97.3%
10	97.6%	97.5%
14	97.7%	98.2%
20	97.9%	97.2%

N = Number of adaptively smoothed spectral vectors produced per digit.

Exp1 = Recognition results obtained using the adaptively smoothed spectral vectors.
Exp2 = Recognition results obtained using the adaptively smoothed spectral vectors augmented with adaptive spectral difference vectors.

Table 1. Recognition results

These results show that adaptively smoothing the estimator to produce as few as 8 spectral vectors per digit has no significant effect on the word recognition accuracy, the observed differences being statistically insignificant. Given that the average digit length was 580ms, this implies that the adaptive smoothing algorithm allows a substantial reduction in the temporal spectral sampling rate, without a significant change in recognition performance.

The second experiment, with the adaptively augmented vector sequences, showed, rather surprisingly, that the spectral difference vectors had no significant effect on recognition accuracy, neither enhancing nor degrading it (apart from when $N = 4$). Although on first glance these results would seem to contradict current thinking, a little reflection shows that this is not necessarily the case. The effect of the spectral difference vectors in a non-adaptively smoothed environment is to weight up the importance of the transition regions in the t-f spectrum. However, the adaptive smoothing scheme already does this implicitly, without the need for difference vectors.

CONCLUSIONS

The advantages of adaptive smoothing schemes are several, including implicit time normalisation of the signal, the effective increased weighting given to transition portions of the spectrum, and a significantly reduced temporal sampling rate. We would also expect the increased average temporal smoothing to improve the signal to noise ratio in the resulting spectral vectors, however this idea has yet to be tested. The major disadvantage of the adaptive smoothing scheme proposed in this paper is the associated increased computational cost. However, this increase is at least partially offset by the reduced computational load incurred by the pattern recognition stage of the recogniser, given the reduced rate at which it need receive the adaptively smoothed vector sequence.

What this paper has failed to demonstrate is any positive increase in recognition performance obtained using the adaptive smoothing strategy. We have shown only that such a strategy enables data rate reduction without incurring a loss of recognition accuracy. However, taking an optimistic view, we would argue that the reason for this may well be the relative ease of the chosen digit recognition task. Consequently, the aim of future work will be to apply the same adaptive smoothing strategy to a speaker independent, large vocabulary recognition task, such as the TIMIT database.

REFERENCES

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

[2]D. Lowe, *Joint Representations in Quantum Mechanics and Signal Processing Theory: Why a Probability Function of Time and Frequency is Disallowed*, Royal Signals and Radar Establishment, 1986, No. 4017

[3]W. Martin and P. Flandrin, *Wigner-Ville Spectral Analysis of Non-stationary Processes*, IEEE Trans. Acoust., Speech, Signal Processing, 1985, pp. 1461-1470

[4]M. D. Riley, *Time-Frequency Representations for Speech Signals*, MIT Artificial Intelligence Laboratory, 1987, AI-TR-974

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

[6]J. Wilbur and F.J. Taylor, *Consistent Speaker Identification via Wigner Smoothing Techniques*, Proc. ICASSP, 1988, pp. 591-594

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

[8]H. Sakoe and S. Chiba, *Dynamic Programming Algorithm Optimisation for Spoken Word Recognition*, IEEE Trans. Acoust., Speech, Signal Processing, 1978, pp. 43-49