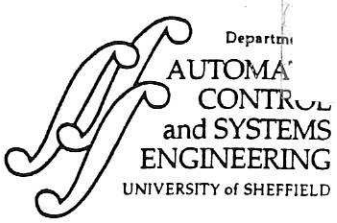


MULTI-DIMENSIONAL CODING OF SPEECH DATA

L Baghai-Ravary [*], S W Beet [**] and M O Tokhi [*]

[*] Department of Automatic Control and Systems Engineering,
[**] Department of Electronic and Electrical Engineering,
The University of Sheffield, Mappin Street, Sheffield, S1 3JD, UK.

Tel: + 44 (0)114 2825136.
Fax: + 44 (0)114 2731 729.
E-mail: O.Tokhi@sheffield.ac.uk.

Research Report No.596

August 1995

Abstract

This paper presents specific new techniques for coding of speech representations and a new general approach to coding for compression, which directly utilises the multi-dimensional nature of the input data. Many methods of speech analysis yield a two-dimensional pattern, with time as one of the dimensions. Various such speech representations, and power spectrum sequences in particular, are shown here to be amenable to two-dimensional compression using specific models which take account of a large part of their structure in both dimensions.

Newly developed techniques, namely, Multi-step Adaptive Flux Interpolation (MAFI) and Multi-step Flow-Based Prediction (MFBP) are presented. These are able to code power spectral density (PSD) sequences of speech more completely and accurately than conventional methods, and at low computational cost. This is due to their ability to model non-stationary, piecewise-continuous, signals, of which speech is a good example.

MAFI and MFBP are first applied in the time domain and then to the encoded data in the second dimension. This approach allows the coding algorithm to exploit redundancy in both dimensions, giving a significant improvement in the overall compression ratio. Furthermore, the compression may be reapplied several times. The data is further compressed with each application.

Key words: Adaptive flux interpolation, flow-based prediction, speech coding.

CONTENTS

LIST OF TABLES AND FIGURES

1 Introduction

The advantage gained in making use of the two-dimensional structure of speech representations' for coding depends on the nature of the pattern and on the availability of an adequate model of the pattern. This paper addresses this problem by adopting approaches based on variable frame-rate (VFR) coding. VFR coding is often advocated as an efficient method for reducing the transmitted data rate in speech communication systems (Holmes, 1974; Papamichalis and Barnwell, 1983; Viswanathan et.al, 1977). In this approach, blocks of a variable number of frames (or vectors) are represented by a single set of parameters. In most cases, these parameters are simply one or more vectors taken from the block. The omitted vectors are then reconstructed (if needed) based on some model of speech dynamics. The sizes of the blocks are chosen to maintain a desired accuracy when the reconstructed vectors are compared to the original vectors, or to maximise the overall accuracy, subject to a constraint on the number of vectors transmitted (Dupree, 1984; Papamichalis and Barnwell, 1983).

VFR coding can also improve the accuracy of automatic speech recognition (ASR) systems by removing redundant observation vectors (Peeling and Ponting, 1989). In any case, the ability to identify and remove the longest possible runs of data is desirable. Current methods are only able to achieve significant compression ratios during quasi-stationary periods. This is because the underlying model used to identify critical vectors of data is too rigid; it only models the changes in values of individual elements within the vectors. They do not identify similar forms of redundancy in the other dimension, or model the more complex forms of redundancy observed in, for example, speech power spectra during non-stationary, but smoothly and predictably evolving, periods.

Techniques which can model such smooth spectral evolution have previously been described (Baghai-Ravary et.al, 1995; Beet et.al, 1994). these, however, have only been applied to estimation of individual vectors of data, not extended sequences of many vectors, as would be required for VFR coding. Furthermore, they have been applied in the temporal direction, and thus have not removed redundancy in the transverse (frequency) direction.

2 Linear models of signal evolution

An autoregressive (AR) linear model can be used to regenerate omitted vectors. This can be generalised as:

$$\mathbf{o}_n = \sum_{i=1}^{N} \mathbf{C}_i(n)\mathbf{o}_{n-i} + \mathbf{v} \qquad (1)$$

where, $\mathbf{o}_n$ is the current parameter vector, $\mathbf{C}_i(n)$ is a (possibly time-dependent) coefficient matrix, $\mathbf{o}_{n-i}$ is the i^{th} preceding observation vector, and $\mathbf{v}_n$ is the n^{th} vector of the innovation (or prediction error) sequence. N determines the order of the model. The model implicit in a zero-order hold system, which would be appropriate for piecewise-constant data, can be obtained by setting $N=1$ and $\mathbf{C}_1(n)=\mathbf{I}$ in equation (1), where $\mathbf{I}$ is the identity matrix[1]. In VFR coding, $\mathbf{v}_n$ is assumed to be zero during each block, and thus the reconstructed vectors are held equal to the first vector of the block. The reconstructed data has a step-wise form, which is not always acceptable. A smoother form of data can be generated with a first-order hold[2], corresponding to $N=2$, $\mathbf{C}_1(n)=2\mathbf{I}$ and $\mathbf{C}_2(n)=-\mathbf{I}$ in equation (1). In this manner, when $\mathbf{v}_n$ is held at zero during VFR blocks, the model gives linear interpolation between the corresponding elements in consecutive transmitted vectors. This is more acceptable, but still imperfect.

While a first-order hold may be appropriate for short quasi-stationary spectral sequences derived from (for example) steady-state vowels, neither of these models can allow for migration of features from one element of the parameter vector to another. In the case of PSDs, this means that changes in frequency cannot be modelled, except by rapidly fading one frequency component out, while fading a new one in.

Some improvement can be obtained by optimising the coefficient matrices, and allowing them to be of arbitrary structure. The $\mathbf{C}_i(n)$ are chosen to match the signal statistics and minimise $\mathrm{E}\left(|\mathbf{v}_n|^2\right)$. Unfortunately, the estimation of values for these matrices is only tractable for the first-order case, and even then it either requires prior knowledge

1 Here, the $\mathbf{v}_n$ sequence is equivalent to the delta coefficients used in many ASR systems.
2 In this case, $\mathbf{v}_n$ forms the delta-delta, or 'acceleration', coefficients used in more sophisticated ASR systems.

about signal statistics[3] or extended periods of pseudo-stationarity in the signal so that they can be estimated locally (Boram, 1990).

3 Non-stationary models

In practice, no AR model based on fixed coefficient matrices can fully simulate the form of observed speech spectrum dynamics. This is illustrated in Figure 1 which shows the spectrogram and corresponding spectrographic of a short segment of speech. Spectrographic flow. Spectrographic flow links the features of successive vectors optimally, with the darkness of the lines representing the value of the respective elements in the preceding vector. As can be seen, in Figure 1, the features within the vectors change position and value with respect to time. The direction of flow of any features through any specific element also changes with time. A fixed AR coefficient matrix, however, will produce a fixed direction of flow for each element. Thus, it would be more realistic if the coefficient matrix would adapt to cope with those changes. Two approaches which are able to handle, for example, steadily changing frequencies and amplitudes in PSD representations are introduced here. These are Flow-Based Prediction (FBP) (Beet et.al, 1994) and Adaptive Flux Interpolation (AFI) (Baghai-Ravary et.al, 1995).

FBP estimates the flow of features from their current positions, forward into the future. Conversely, AFI identifies optimal links (termed lines of flux) between the features of the immediate neighbours of the unknown vectors and interpolates along them. However, both methods assume that the positions and values of any features within the parameter vector change linearly with time. This corresponds to $\mathbf{C}_i(n)$ matrices which evolve during each block. The evolution, however, is systematic and defined only by a pair of vectors. Both methods fit into the framework described below:

A block of M observation vectors, $\mathbf{o}_n$ to $\mathbf{o}_{n+M-1}$, can be estimated by prior knowledge of two of those vectors, $\mathbf{o}_\alpha$ and $\mathbf{o}_\beta$, by assuming that (within the block) links

---

3 In ASR applications this training can be achieved by tying the model's coefficient matrices to individual states within a hidden Markov model (HMM), to give a linear predictive HMM (Kenny et.al, 1990).

exist, joining similar values in successive vectors. In the methods described here, it is assumed that these lines are straight, and that the values of the parameter vectors evolve linearly along them. Mathematically, if a link passes through element i of $\mathbf{o}_\alpha$ and element j of $\mathbf{o}_\beta$, then the corresponding element, k, of $\mathbf{o}_\gamma$ will be given by

$$\begin{aligned} o_{\gamma,k} &= o_{\alpha,i} + \frac{(\gamma-\alpha)}{(\beta-\alpha)}\left(o_{\beta,j} - o_{\alpha,i}\right) \\ k &= \frac{(\gamma-\alpha)}{(\beta-\alpha)}(j-i) \end{aligned} \tag{2}$$

If $\mathbf{o}_\alpha$ and $\mathbf{o}_\beta$ precede $\mathbf{o}_\gamma$ the model is essentially a predictor. In particular, in the special case when $\gamma = \beta + 1 = \alpha + 2$, one-step-ahead prediction is performed, as in FBP. If $\mathbf{o}_\gamma$ lies between $\mathbf{o}_\alpha$ and $\mathbf{o}_\beta$, it is an interpolator and if $\beta = \gamma + 1 = \alpha + 2$, one-step interpolation results, as in AFI. In both these (one-step) special cases, the set of non-crossing links which minimise $|\mathbf{v}_n|^2$ can be found efficiently using dynamic programming, yielding estimates of the $\mathbf{C}_1(n)$ and $\mathbf{C}_2(n)$ matrices of equation (1).

4 Multi-step estimation

Unfortunately, if the known vectors $\mathbf{o}_\alpha$, $\mathbf{o}_\beta$ and the unknown $\mathbf{o}_\gamma$ are separated by more than one vector, a paradox becomes apparent; if links are non-parallel, they will cross at some point. If all links have to pass exactly through the elements of all three vectors, the crossing will often happen within the block, causing a re-ordering of features which is not normally appropriate. Furthermore, some elements of the vectors to be estimated can be left 'hanging', that is, with no links passing through them. It is, therefore, not possible to maintain straight links for extended periods if they are constrained to intercept all the elements of all the vectors they cut. This is illustrated in Figure 2, where the size of the dots represent the values of the corresponding elements of the (vertical) vectors of data and the links have been calculated between the two known vectors and extrapolated to the end of the block. Where links merge, the data value has been taken to be the average of the individually predicted values. It can be seen in Figure 2 that a single peak in the original (known) vectors has evolved into two separate peaks. This is unlikely to have

occurred in a real observation sequence. Furthermore, the third element from the bottom of the Figure has become detached and would require additional algorithms to allow its value to be estimated.

The solution proposed here involves two parts. Firstly, by always estimating the links between the vector to be estimated and one of its immediate neighbours, no elements of the vector being estimated will be bypassed by the links. Thus, no elements will be undefined. Secondly, by only ever performing 'one-step' estimation, any effects due to crossovers of features are minimised. To allow a one-step estimator to perform multi-step estimation, it is applied recursively; each new estimated vector is used as one of the known vectors of a subsequent estimation. This is shown in Figure 3, with which the problems illustrated in Figure 2 can be seen to have been avoided completely. In this example the peak in the known vectors rapidly dies out, but the flow of features throughout the block is always intuitively reasonable, and no implausible predictions are generated. This principle can be applied in a number of different ways, depending whether an interpolative or predictive approach is adopted. Some of the available options are discussed in the following section.

4.1 Recursive estimation algorithms

In these methods, the links between the two adjacent vectors (one of which is the unknown vector to be estimated) are restricted to avoid cross-overs and to remain within the bounds of the available data, that is links which would extend beyond the ends of the vectors are disallowed.

Recursive MFBP estimates $\mathbf{o}_\gamma$ using the two preceding observation vectors. Thus, $\gamma = \beta + 1 = \alpha + 2$, as described earlier. If the block starts with $\mathbf{o}_n$, then $\mathbf{o}_n$ and $\mathbf{o}_{n+1}$ are used to form the first estimate, $\mathbf{o}_{n+2}$, which is then used with the true observation vector, $\mathbf{o}_{n+1}$, to calculate $\mathbf{o}_{n+3}$. This recursive estimation is then continued so that each observation vector is predicted by the previous two estimates until $\mathbf{o}_{n+M-1}$ is found from the predictions $\mathbf{o}_{n+M-2}$ and $\mathbf{o}_{n+M-3}$. Using this approach, two vectors are transmitted to encode each complete block. The process is illustrated in Table 1 and in Figure 4. In

Figures 4 to 8, dark colour indicates transmission, light colour, estimation, and the arrows originate from the input vectors of the estimator and point to the respective output.

A saving on transmitted data can be achieved by using the last estimated vector of the previous block as if it were the first vector of the block currently under consideration. This is illustrated in Table 2 and in Figure 5. In this manner, the number of vectors needed to describe each block can be halved, since only the first vector of each block need be transmitted. However, any error which accrue from the previous block will carry over and affect the quality of the current block's estimates.

Recursive MAFI also transmits only one vector per block, but errors do not carry over from block to block in the same way as for the 'single-vector' recursive MFBP. Recursive MAFI estimates both $\mathbf{o}_{n+1}$ and $\mathbf{o}_{n+M-2}$ simultaneously using observation vectors $\mathbf{o}_n$ and $\mathbf{o}_{n+M-1}$. The estimates, $\mathbf{o}_{n+1}$ and $\mathbf{o}_{n+M-2}$, are then used to find the next two vectors $\mathbf{o}_{n+2}$ and $\mathbf{o}_{n+M-3}$ and so on, until the block is completed (see Table 3 and Figure 6.).

All the recursive approaches ensure that the assumptions of the MAFI and MFBP models are maintained locally.

4.2 Non-recursive estimation algorithms

Recursive methods require repeated computation of local distance matrices within each block, which would not be necessary if equation (2) could be applied directly. Although the considerations outlined earlier indicate that a recursive approach is necessary, a close approximation can be made by considering the links between an unknown vector and its immediate neighbour (whether known or unknown). The links are assumed to extend to the known vectors according to equation (2), which are then used to estimate the values of the unknown elements. Any implied reordering of features or omission of links through elements in other vectors are ignored. Figures 7 and 8 show the equation of non-recursive MFBP and MAFI, respectively. Since the links are calculated separately for each unknown vector, there is no enforced consistency between the links used to estimate the values of

consecutive vectors. In practice though, this is rarely noticeable in terms of its effect on the estimated values of the vectors.

As in the case of the recursive algorithms described earlier, non-recursive MFBP can be implemented in either one or two-vector per block versions. However, as the performance of recursive and non-recursive approaches are so similar (shown later), only the single-vector version is discussed here. Non-recursive MAFI also operates in a similar way to recursive MAFI, with each known vector being used in reconstruction of pairs of consecutive blocks.

4.3 VFR coding

In the experiments described here, the magnitude of error in an estimation is assessed by a normalised form of mean square error for each vector. This is defined as

$$\xi = \frac{\text{error power}}{\text{signal variance} + \text{noise power}} \times 100\% \tag{3}$$

VFR encoding is then implemented by incrementing the estimation block length from one vector upwards, until the peak estimation error exceeds a threshold, ξ_{max}. At this point the last pair of known vectors to have satisfied the error criterion, are transmitted (unless, as in the MAFI cases, one of them has already been sent) and the system is reinitialised. Along with the transmitted vector(s), an integer is transmitted to indicate the size of the respective block. The decoder then applies the same form of interpolation to regenerate estimates of the omitted vectors.

5 Two-dimensional coding

Having encoded the speech representations by removing redundant vectors, further compression can be achieved by, for example, removing redundant frequency slices (in the case of spectrographic representations). This is to say that the two-dimensional pattern representing the speech can be encoded, transposed, and compressed further by the same algorithm (see Figure 9). In the case of PSD representations of speech, the re-coding in the

transverse direction removes frequency regions which can be easily estimated from their neighbours. The use of MAFI/MFBP estimation allows this estimation to occur even in frequency bands which contain temporal changes of frequency (for example rising and falling formants). Having removed redundant frequency slices, it is sometimes possible to remove more temporal slices, so that the whole process can be repeated *ad infinitum.*

6 Speech representations

To evaluate the capability of MAFI and MFBP, a number of experiments have been performed using a diverse range of speech representations. The parameters and methods used to generate these representations were chosen to be consistent to give valid comparisons (Baghai-Ravary et.al, 1994). Most of these analysis methods have auditory counterparts, where the frequency resolution and the frequency and loudness scales of the human auditory system are applied. These were also implemented. The methods described here include the pre-processors described below.

6.1 Periodogram

This is the most common method for visualising speech signals. It is formed by taking the DFT of a windowed segment of speech, and finding the modulus squared of each complex output value;

$$\text{Discrete Fourier Transform} = X(m) = \sum_{n=0}^{N-1} x(n)h(n)\exp\left(-j2\pi\frac{nm}{N}\right)$$

$$\text{Periodogram} = |X(m)|^2 \qquad (4)$$

In this manner, it provides an estimate of the PSD which is only degraded by the spectral effects of the temporal window, $h(n)$. The frequency resolution of the periodogram is inversely proportional to the length of the input frame (for a given window shape), and cannot be controlled independently.

The choice of window is restricted by the expected dynamic range of the elements in each PSD estimate, and the required degree of temporal continuity. To give temporal

continuity across pitch-pulses, with adult male speech, this method can only give a narrow-band spectrogram, clearly resolving individual pitch harmonics. Such data has a complicated flow structure, since the frequency of the pitch harmonics may, for example, rise, while the formants are falling.

As with most of the representations presented here, the most obvious difference between the periodogram and its auditory version, is the frequency warping; at low frequencies, the auditory representation is stretched, while at high frequencies, it is severely compressed. In this case, however, the changing bandwidth of the auditory representation has a further effect on the representation. At low frequencies, where the auditory bandwidths are small, individual pitch harmonics are clearly resolved. However, even during high-pitched segments, the pitch harmonics are suppressed at high frequencies.

6.2 Blackman-Tukey power spectrum

One method for controlling the resolution of a periodogram is to window an estimate of the autocorrelation function, rather than the data itself. This allows the frequency resolution to be reduced without losing temporal continuity. However, the window must have a non-negative Fourier transform for negative PSD estimates to be avoided. In the work reported here, the window was made equal to the autocorrelation function of a suitable prototype window, simultaneously ensuring that the created window has finite duration and a real, and non-negative, Fourier transform.

$$\begin{aligned} \text{Autocorrelation function} &= A_x(m) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)x(n+m) \\ \text{Window} &= A_h(m) = \frac{1}{N}\sum_{n=0}^{N-1} h(n)h(n+m) \\ \text{Blackman - Tukey PSD estimate} &= \text{DFT}\left(A_x(m)A_h(m)\right) \end{aligned} \tag{5}$$

This method can thus give a broad-band spectrogram, characterising formant structure rather than pitch. The particular window used here gave a spectral resolution of 400Hz, to ensure that even the highest female pitch was suppressed. However, since the resonances

of the vocal tract are generally narrower than this (Deller et.al, 1993), the resulting spectrum is somewhat blurred. Furthermore, the change in frequency resolution due to the auditory transformation is minimal; most auditory filters have bandwidths less than the 400Hz 'blurring' introduced by the autocorrelation window.

6.3 *Maximum entropy power spectrum*

The power spectrum of an AR process can be obtained by calculating the parameters of the AR model from the autocorrelation function of the signal. This has been done here using Burg's method. The maximum entropy method (MEM) PSD estimate is then obtained by multiplying the innovation power by the transfer function of the implied recursive filter;

$$\begin{aligned} \text{Innovation power} &= \sigma^2 \\ \text{AR transfer function} &= H_{AR}(F) = \frac{1}{1+\sum_{m=1}^{p} a_m \exp\left(-j2\pi \frac{F}{F_s} m\right)} \\ \text{Maximum Entropy PSD} &= \sigma^2 \left|H_{AR}(F)\right|^2 \end{aligned} \tag{6}$$

where, F_S is the sampling frequency. Since speech can only be loosely approximated as an AR process, the resulting PSD estimate tends to be "peaky", with occasional false peaks (so-called "peak splitting"). Nonetheless, it can make formant tracks very clear, both in the conventional and the auditory versions.

6.4 *Maximum likelihood power spectrum*

This is variously referred to as the minimum variance PSD estimate, the maximum likelihood method (MLM) or Capon's method. It involves the design of an FIR filter for each frequency where an estimate of the PSD is required. These filters have unity gain at the design frequency, but are designed to give minimal output power. The technique attempts to attenuate all but the frequency component of interest, and can be considered as a data-adaptive periodogram.

The order of the filters determines the maximum number of frequency regions which can be attenuated, and is chosen according to the application. To resolve formant structure while suppressing pitch information, the filter order should be chosen so as to be slightly more than twice the maximum number of formants, as in linear prediction analysis.

The power from each filter is calculated from the AR model for the signal, without explicitly implementing the filters (Musicus, 1985);

$$\text{Maximum Likelihood PSD} = \frac{\sigma^2}{\sum_{k=-p}^{p} \mu(k) \exp\left(j2\pi \frac{F}{F_s} k \right)}$$

$$\text{where } \mu(k) = \sum_{m=0}^{p-k} (N + 1 - |k| - 2i) a_m a_{m+|k|} \text{ and } a_0 = 1 \tag{7}$$

The MLM method generally has frequency resolution comparable with, or slightly better than, that of the Blackman-Tukey method, and has superior temporal continuity. The latter effect is due to the Blackman-Tukey method's need for data windowing, which is avoided in the maximum likelihood approach, being based (in this example) on Burg's method of AR parameter estimation.

6.5 Cepstrum

Since speech can be considered as the product of a source spectrum and a vocal tract transfer function, pitch information can be separated from formant structure by homomorphic filtering. A log-power periodogram is formed and then inverse-Fourier transformed to give a cepstrum containing formant data in its lower coefficients, with pitch being apparent at the higher end;

$$\text{Cepstrum} = \text{DFT}^{-1}\left(\ln \left| \text{DFT}(x(n)h(n)) \right|^2 \right) \tag{8}$$

In practice, it is computationally more efficient to use an inverse discrete cosine transform (DCT) rather than an inverse DFT, since the periodogram is, by definition, real and, if the original speech signal is real, symmetric about zero frequency.

The first cepstral coefficient is equivalent to the log signal power, and thus has a wider range than the rest. The remaining coefficients describe the overall spectral shape. Thus, all but the first coefficient tend to change smoothly and with consistent 'flow'. The higher coefficients (which characterise pitch) evolve especially smoothly in the conventional cepstrum, but are almost non-existent in the auditory version (because the pitch harmonics are rendered aperiodic by the non-linear frequency warping).

6.6 Linear prediction coefficients

Autoregressive modelling of speech signals can give a very concise description of the vocal tract transfer function. The results of this analysis are often presented as coefficients of a ladder filter which can be used to predict one step ahead of the speech waveform. However, they can exhibit abrupt changes even when the speech sound is changing smoothly. It is difficult to identify meaningful structure in them, and they only appear smooth during unvoiced sounds. There is little evidence of migration of features between one coefficient and another.

6.7 Reflection coefficients

Burg's method for calculating linear prediction coefficients is based on the calculation of reflection coefficients, which can be viewed as parameters of an acoustic-pipe model of speech production (Rabiner and Schafer, 1978). These always have values between -1 and 1, and thus have somewhat different numerical properties from those of standard linear prediction (ladder) coefficients. They also tend to evolve slightly more smoothly than ladder coefficients, and do exhibit slight migration of features.

6.8 Vocal tract area functions

The shape of the acoustic pipe implied by a set of reflection coefficients can be calculated by adding successive log area ratios or alternatively, by multiplying the area ratios and then taking the log;

$$\text{Area of section i} = A_i = A_{i-1}\frac{1-k_i}{1+k_i} \quad ; \quad 1 \le i \le p,\ A_0 = 1$$

$$\text{Log area of section i} = \ln\left(\prod_{m=1}^{i}\frac{1-k_m}{1+k_m}\right) \tag{9}$$

This gives a set of parameters which are loosely related to the cross-sectional area of the vocal tract, and therefore obey rules of motion similar to those of the real vocal tract. For example, as the tongue moves a constriction forward and backward, values of the vocal tract area function will move within the data vector, while the opening and closing of the mouth will affect the magnitude of the values at the respective end of that vector. This type of representation is much smoother than either reflection coefficients or linear prediction coefficients, and exhibits noticeable migration effects.

6.9 Autocorrelation function

The preceding methods are closely related to the autocorrelation function (ACF) of the speech signal. Therefore, it seems logical to evaluate this function as well. The most efficient method of estimating the ACF is to take the inverse Fourier transform of the periodogram (or the inverse DCT of the positive-frequency part of the periodogram). This results in an ACF estimate which is distorted by the initial windowing of the speech, but exhibits good temporal continuity.

7 Results

The different forms of MAFI and MFBP variable frame-rate coding described in sections 4 and 5 were implemented and mean compression ratios computed for all the pre-processors discussed in section 6. A summary of the results are shown in Figures 10 to 13. These are discussed below.

7.1 Multi-step estimators for coding

For the results described in this section, the maximum likelihood method was used to generate broad-band spectrograms. This form of data is well-suited to the assumptions of

MAFI and MFBP, but gives broadly comparable trends to those observed for other forms of data.

Figure 10 shows the variation in the compression ratio of each coding method's with the transmission threshold, ξ_{max}, used to specify the maximum allowable distortion. To illustrate the level of distortion implied by this threshold, Figures 11(a)-(c) show a typical spectrogram of a spoken sentence, the reconstructed version after one application of recursive MAFI, and the error magnitude of the reconstruction, respectively, for a transmission threshold of 20%. Thus, a threshold of 20% corresponds to a barely-noticeable coding distortion, but, even so, it can give a compression ratio of over 50%. Thus half of the original data vectors are redundant. Increasing the threshold consistently increases the compression ratio, albeit at the expense of reduced accuracy. However, the most significant observation concerning these results is that recursive and non-recursive MAFI perform almost identically at all levels, while the MFBP approaches give roughly 30% lower compression ratios.

7.2 Multi-dimensional coding

Again, these results concentrate on the maximum likelihood broad-band spectrogram, but the observations made here are also applicable to other representations, to varying degrees. Figure 12 shows the compression ratios achieved after varying numbers of applications of MAFI and MFBP in orthogonal directions. In all cases, ξ_{max} was set to 20% to give a small coding error. It is clear that significant redundancy is still present after one application of these coding methods; the second application of each method generally increases the compression ratio by a further 30 to 40%. An example of the quality of the coding after two applications is shown in Figures 11(d)-(e), where the reconstructed version of Figure 11(a) and the error magnitude are shown, respectively. Beyond two applications, the compression continues to increase until, after about 6 applications, little further compression is obtained. It is worth noting that although the difference between the MAFI methods and MFBP is around 30% after one application, this has dropped

significantly after two applications. However, MAFI is still noticeably better than MFBP in all cases.

7.3 Representations

As mentioned above, MAFI yields greater compression ratios for a particular transmission threshold, ξ_{max}, than MFBP, and there is negligible difference between the recursive and non-recursive forms. Therefore, for these experiments, recursive MAFI was selected and applied repeatedly (in orthogonal directions, as described in section 5) to all the speech representations of section 6.

In all cases, subsequent applications of the MAFI algorithm gave increased compression of the various speech representations, as summarised in Figure 13. However, the additional compression yielded by subsequent applications varies considerably, depending on the representation. In particular, the 'time-domain' representations, namely, the cepstrum, autocorrelation function, linear prediction coefficients and reflection coefficients, are only noticeably compressed in the temporal dimension. All the other representations are compressed by broadly comparable amounts in each dimension.

In general, compression levels of around 80 to 90% are achieved with most representations after two applications. The exceptions to this are the 'time-domain' techniques mentioned before, and the periodogram, which exhibits a more complex structure than the others, with different aspects of the spectral features often moving in different directions (that is, pitch harmonics can rise while formants are falling, or *vice versa*). It is of interest to note that the auditory periodogram does not suffer from this problem to anything like the same extent; at high frequencies, its resolution is restricted by that of the human auditory system, while at low frequencies, the frequency scale is expanded, making the spectral features broader, and thus easier to predict (in the frequency-domain).

8 Conclusion

MAFI and MFBP have been presented as new methods for compressing various representations of speech. Variants of each of these methods have been outlined and evaluated by comparative experiments. These have shown that both recursive and non-recursive methods of implementing MAFI give virtually identical performances. This implies that the non-recursive form would be preferred since it is computationally more efficient. However, it has been observed informally that any coding artefacts introduced by the recursive form have been less obtrusive than those resulting from the non-recursive algorithm. Thus, the choice might not always be so straightforward.

It has also been shown that, as expected, MAFI out-performs MFBP, since interpolation is inherently more robust than prediction. Surprisingly, however, there is little difference in the compression ratios of the single-frame and two-frame variants of recursive MFBP. This implies that the single-frame 'error carry-over' effect can be quite severe, so that the average block length is halved relative to the two-frame case.

As far as the issue of dimensionality is concerned, it is clear from the results presented here that all representations, except those with both dimensions representing time, benefit from a multi-dimensional approach to coding.

9 References

BAGHAI-RAVARY, L., BEET, S. W. and TOKHI, M. O. (1994). Removing redundancy from some common representations of speech, *Proceedings of the Institute of Acoustics*, **16**, (Part 5), pp. 467-74.

BAGHAI-RAVARY, L., BEET, S. W. and TOKHI, M. O. (1995). Adaptive flux interpolation, flow-based prediction, delta or delta-delta coefficients: which is best?, *Proceedings o. Eurospeech '95*, Madrid, September 1995. (To appear).

BEET, S. W., BAGHAI-RAVARY, L. and TOKHI, M. O. (1994). Non-stationary prediction of frame-based speech data", in M. Holt, C. Cowan, P. Grant and W.

Sandham (editors) *Signal Processing VII: Theories and applications*, EURASIP, Lausanne, Switzerland, pp 1649-52.

BORAM, Y. (1990). Construction of linear predictors for stationary vector sequences, *IEEE Trans. on Automatic Control*, **35**, (2), pp. 236-9.

DELLER, J. R. Jr., PROAKIS, J. G. and HANSEN, J. H. L. (1993). *Discrete-Time Processing of Speech Signals*, Macmillan, New York.

DUPREE, B. C. (1984). Formant coding of speech using dynamic programming, *Electronics Letters*, **20**, (7), pp. 279-280.

HOLMES, J. N. (1974). A variable-frame-rate coding scheme for speech analysis-synthesis systems, *Electronics Letters*, **10**, pp. 101-102.

KENNY, P., LENNIG, M. and MERMELSTEIN, P. (1990). A Linear Predictive HMM for Vector-Valued Observations with Applications to Speech Recognition, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **38**, (2), pp. 220-225.

MUSICUS, B. R. (1985). Fast MLM power spectrum estimation from uniformly spaced correlations, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **33**, pp. 1333-1335.

PAPAMICHALIS, P. E. and BARNWELL, T. P. III (1983). Variable rate speech compression by encoding subsets of PARCOR coefficients, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **31**, (3), pp. 706-13.

PEELING, S. M. and PONTING, K. M. (1989). Experiments in variable frame-rate analysis for speech recognition, *RSRE Research Memorandum 4330*, Royal Signals & Radar Establishment, Malvern, UK.

RABINER, L. R. and SCHAFER, R. W. (1978). *Digital Processing of Speech Signals*, Prentice Hall, Englewood Cliffs.

VISWANATHAN, R. *et al.* (1977). The application of a functional perceptual model of speech to variable-rate LPC systems, *Proceedings of IEEE ICASSP-77*, pp. 219-222.

Table 1: Predictor inputs and transmitted data (shaded) for 'two-vector' recursive MFBP

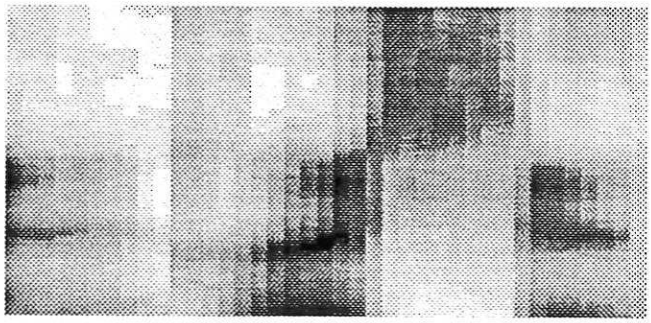
First input o_α	Second input o_β	Prediction o_γ
⋮	⋮	⋮
$\hat{o}_{n-3}$	$\hat{o}_{n-2}$	$\hat{o}_{n-1}$
—	—	o_n
—	—	o_{n+1}
o_n	o_{n+1}	$\hat{o}_{n+2}$
o_{n+1}	$\hat{o}_{n+2}$	$\hat{o}_{n+3}$
⋮	⋮	⋮
$\hat{o}_{n+M-3}$	$\hat{o}_{n+M-2}$	$\hat{o}_{n+M-1}$
—	—	o_{n+M}
—	—	o_{n+M+1}
⋮	⋮	⋮

Table 2: Predictor inputs and transmitted data (shaded) for 'one-vector' recursive MFBP

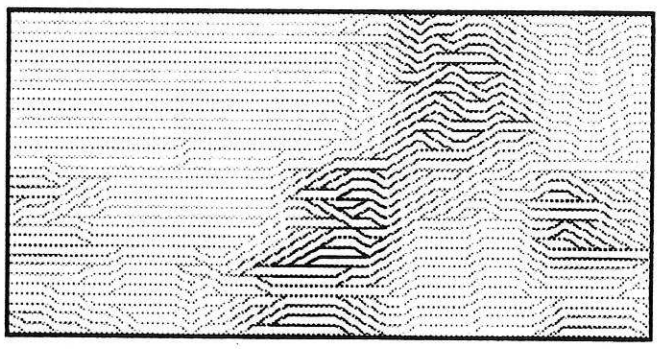
First input o_α	Second input o_β	Prediction o_γ
⋮	⋮	⋮
$\hat{o}_{n-3}$	$\hat{o}_{n-2}$	$\hat{o}_{n-1}$
—	—	o_n
$\hat{o}_{n-1}$	o_n	$\hat{o}_{n+1}$
o_n	$\hat{o}_{n+1}$	$\hat{o}_{n+2}$
$\hat{o}_{n+1}$	$\hat{o}_{n+2}$	$\hat{o}_{n+3}$
⋮	⋮	⋮
$\hat{o}_{n+M-3}$	$\hat{o}_{n+M-2}$	$\hat{o}_{n+M-1}$
—	—	o_{n+M}
⋮	⋮	⋮

Table 3: Predictor inputs and transmitted data (shaded) for recursive MAFI.

First input o_α	Second input o_β	Prediction o_γ
⋮	⋮	⋮
$\hat{o}_{n-3}$	$\hat{o}_{n-2}$	$\hat{o}_{n-1}$
—	—	o_n
o_n	o_{n+M}	$\hat{o}_{n+1}$
$\hat{o}_{n+1}$	$\hat{o}_{n+M-1}$	$\hat{o}_{n+2}$
$\hat{o}_{n+2}$	$\hat{o}_{n+M-2}$	$\hat{o}_{n+3}$
⋮	⋮	⋮
$\hat{o}_{n+2}$	$\hat{o}_{n+M-2}$	$\hat{o}_{n+M-3}$
$\hat{o}_{n+1}$	$\hat{o}_{n+M-1}$	$\hat{o}_{n+M-2}$
o_n	o_{n+M}	$\hat{o}_{n+M-1}$
—	—	o_{n+M}
⋮	⋮	⋮



(a)



(b)

Figure 1: A short segment of speech;
(a) Maximum likelihood spectrogram.
(b) Spectrographic flow.

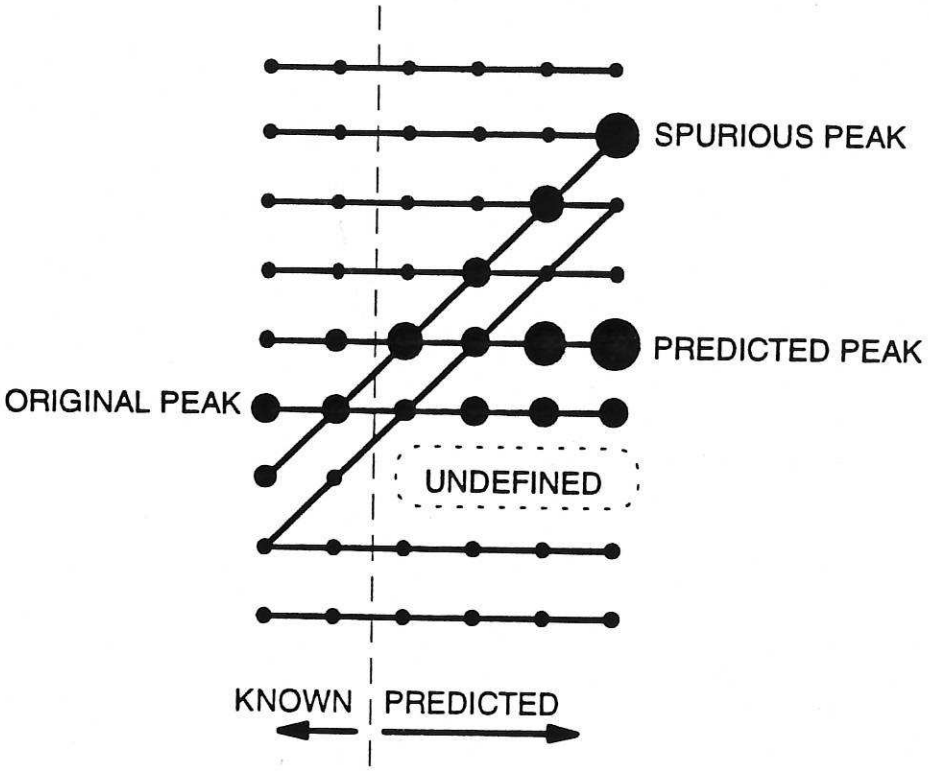


Figure 2: An example of the problems caused by inappropriate application of the basic FBP model.

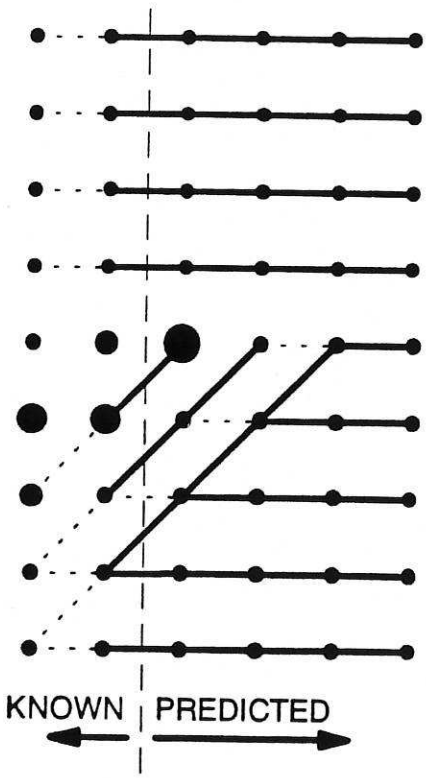


Figure 3: Avoidance of the problems illustrated in Figure 2 by recursive 'one-step' estimation (links implied, but not explicitly calculated, are shown dotted).

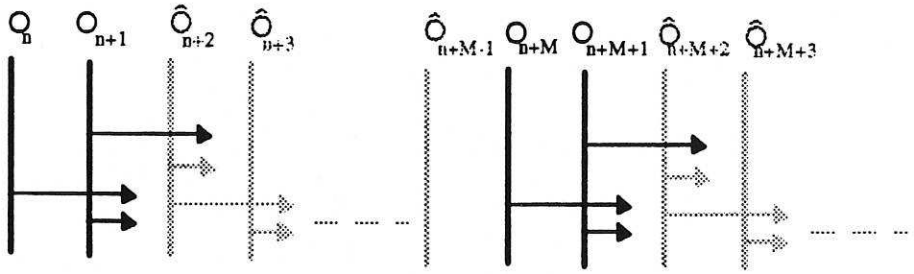


Figure 4: Two-vector recursive MFBP process.

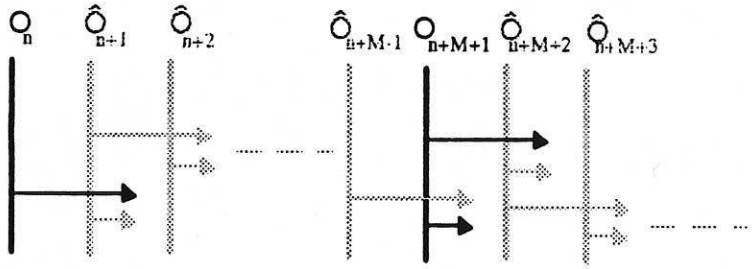


Figure 5: One-vector recursive MFBP process.

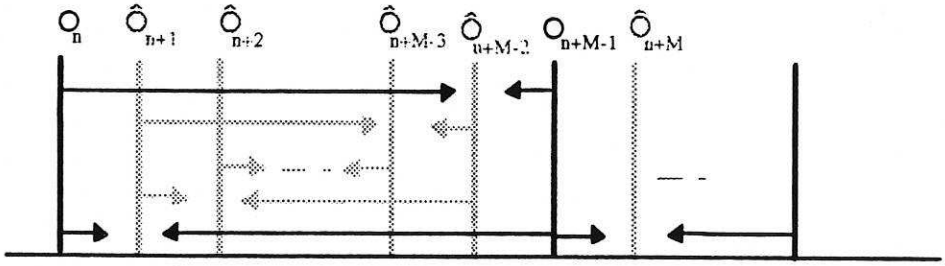


Figure 6: Recursive MAFI process.

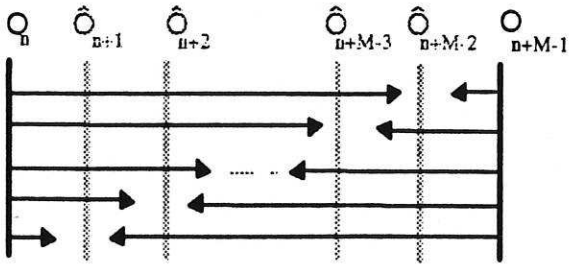


Figure 7: Non-recursive MAFI process.

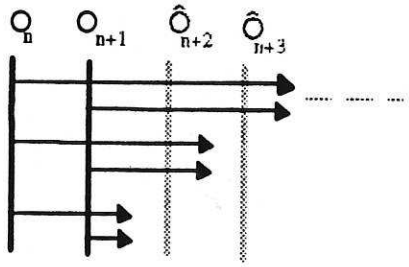


Figure 8: Non-recursive MFBP process.

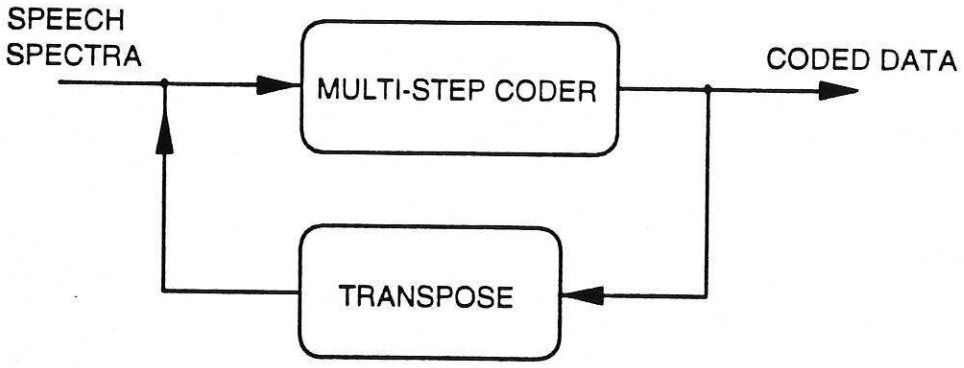


Figure 9: Procedure for 2-dimensional coding.

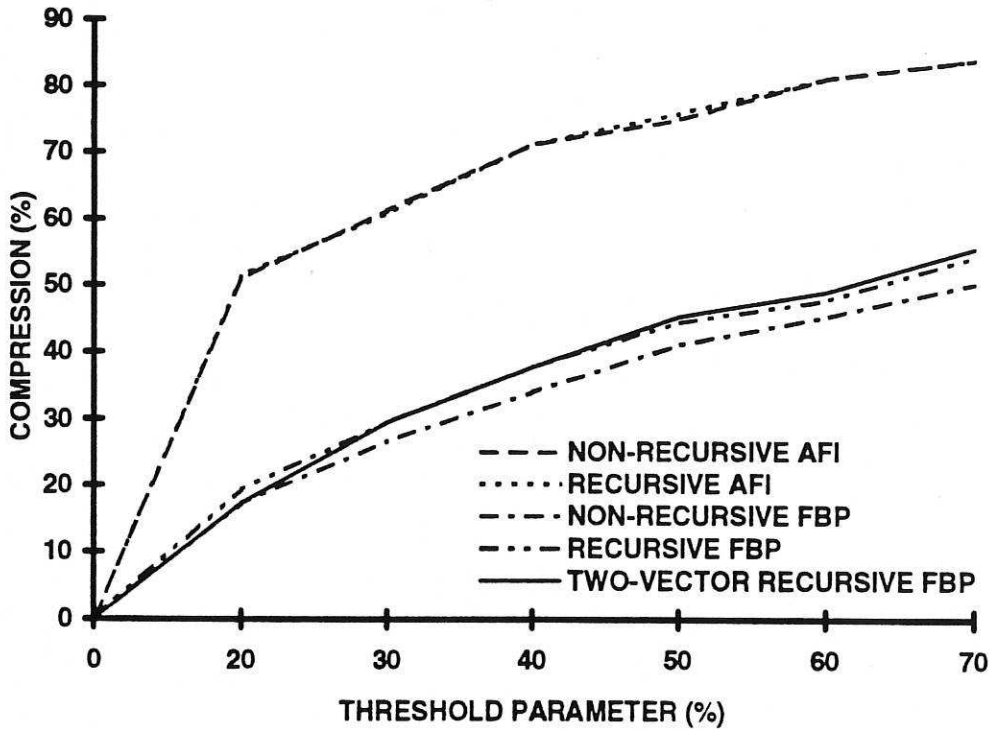
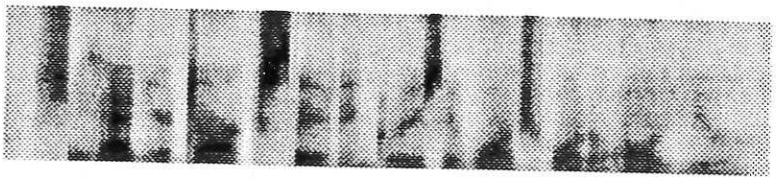
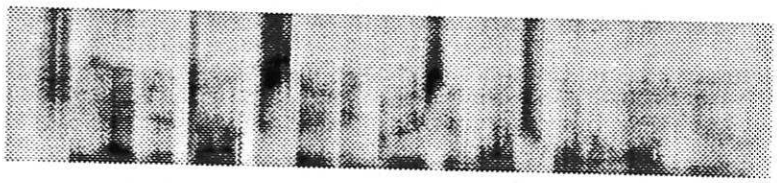


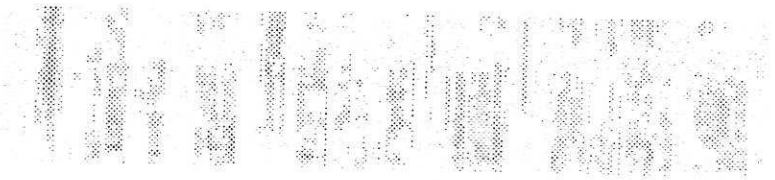
Figure 10: Maximum likelihood spectrogram compression as a function of threshold, ξ_{max}, after one application of recursive MAFI.



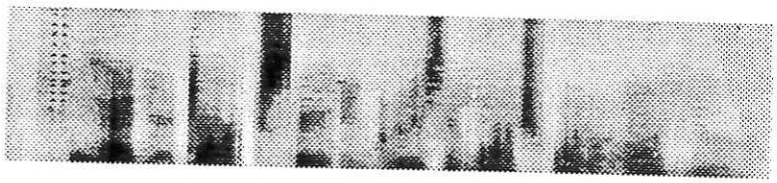
(a)



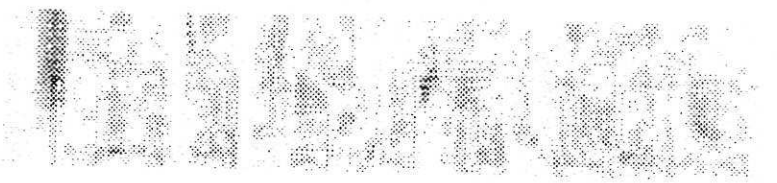
(b)



(c)



(d)



(e)

Figure 11: (a) Maximum likelihood spectrogram of a sentence spoken by a female.
(b) The reconstructed spectrogram after one application of MAFI.
(c) The corresponding reconstruction error.
(d) The reconstruction after 2 applications.
(e) The error corresponding to (d).

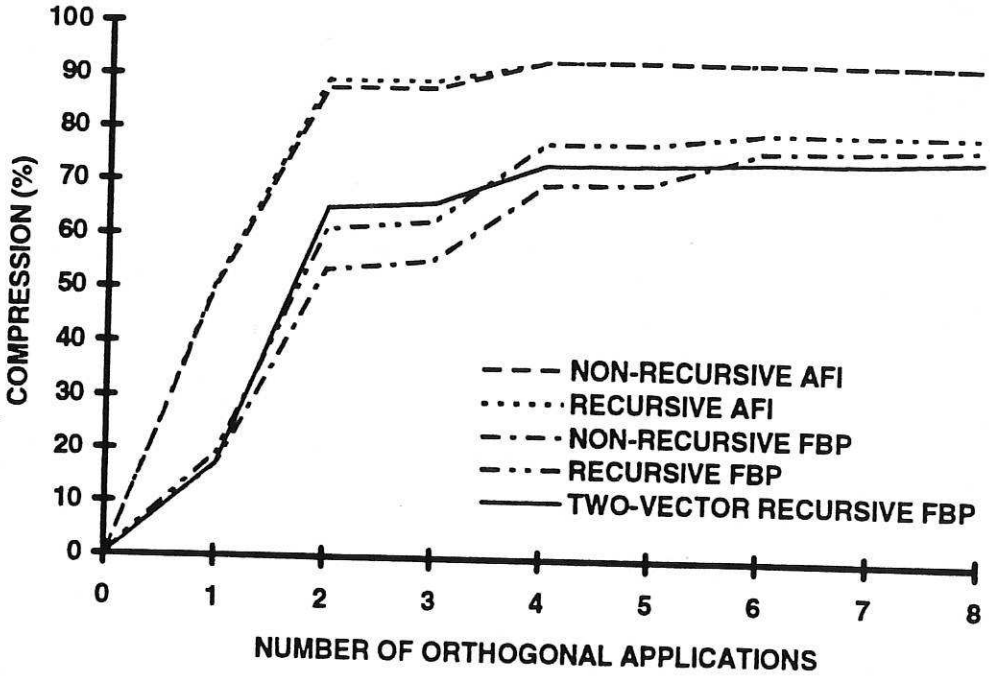


Figure 12: Maximum likelihood spectrogram compression as a function of number of recursive MAFI application for $\xi_{max} = 20\%$.

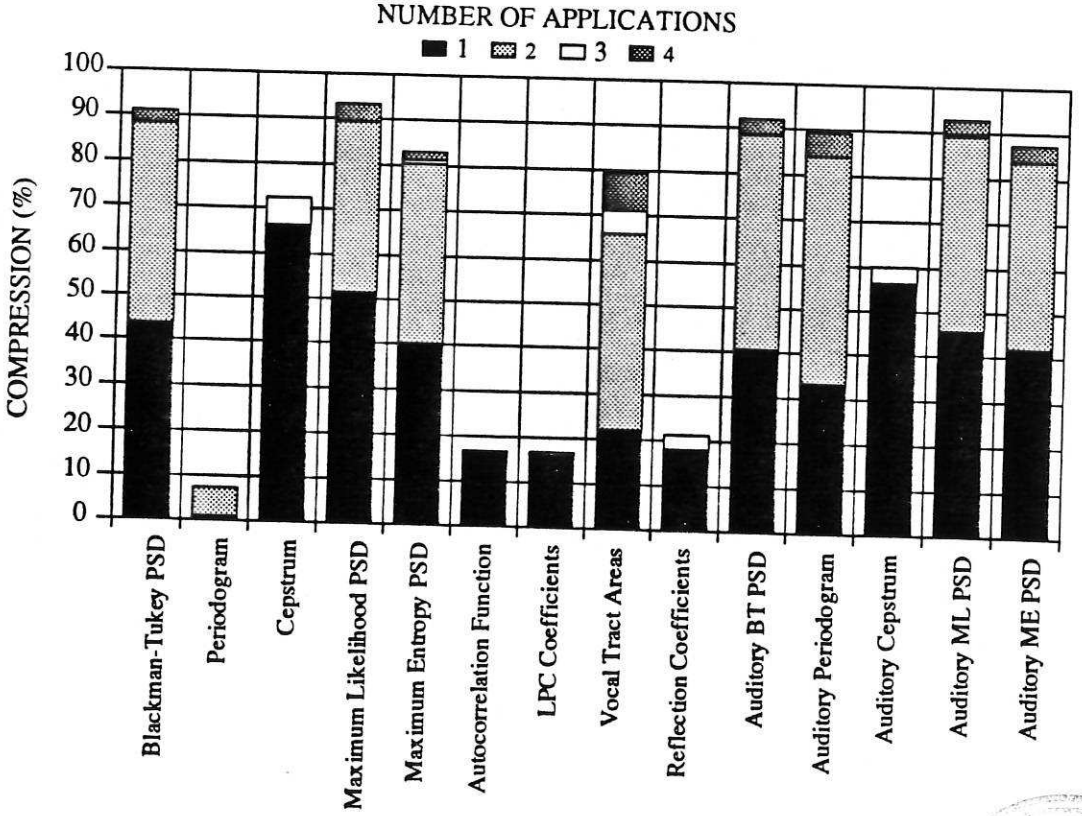


Figure 13: Compression ratios achieved for various preprocessors.