THE TWO-DIMENSIONAL DISCRETE COSINE TRANSFORM APPLIED TO 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.

ABSTRACT

A two-dimensional discrete cosine transform (2-d DCT), often used for image coding, has been applied to sequences of speech spectra produced by the maximum likelihood method (MLM).

The coded data was compressed by over 90%, reducing it to a size smaller than that needed to store the coefficients of a 10^{th} order linear predictive coding (LPC) model. The DCT-encoded data was then reconstructed and tested for intelligibility. It was found that the two-dimensional DCT method was significantly more intelligible and more natural than the LPC technique.

Moreover, when the power spectrum data is augmented with its phase information prior to compression with the two-dimensional DCT, a further reduction in the volume of data required for synthesis is obtained.

INTRODUCTION

In the past, parametric methods, such as linear prediction, have been widely accepted as the only practical approach to very low data rate speech coding (e.g. [1], page 489). Their saving in volume of data is due to two important factors:

1. an innovation signal (to drive a simulated vocal tract filter) which can be adequately characterised by only two parameters (pitch and intensity).
2. a small number of coefficients (model parameters) needed to describe the filter driven by the innovation signal.

The intelligibility and quality of the reconstructed speech will depend on the ability of the filter to match the true spectral shape of the signal, and on the synthetic innovation signal to reproduce an acceptable phase characteristic and harmonic structure. The use of pitch-synchronous analysis techniques can minimise the need to explicitly model large-scale temporal factors.

Problems:

The all-pole autoregressive (AR) model implicit in LPC is quite well-suited to the modelling of non-nasalised speech in quiet surroundings and with a known acoustic transfer function between the speaker and the microphone. However, nasalised sounds, and noisy and unpredictable acoustic environments can all degrade quality significantly.

A more robust system would be to transmit parameters directly related to perception of speech (since variability in the input signal is relatively large). To date, this approach has not been so widely exploited because of the difficulty of defining a set of parameters

which describe perceptual phenomena adequately, while being compact enough to allow low data-rate coding.

A clue to the solution of this problem lies in that auditory perception is essentially a two-dimensional process (for example, there are contrast-enhancing mechanisms within the peripheral auditory system, which operate in both the time and the frequency domains).

A Solution:

Although it is difficult to identify a small number of parameters which adequately describe the perception of a single short segment (frame) of speech, there is significant correlation between successive frames. Thus a sequence of many frames can be encoded with little more data than a single frame.

Optimal coding could be performed using a Karhunen-Loève transform (KLT) of each block, but as with image data, a close approximation can be achieved with a two-dimensional discrete cosine transform (2-d DCT). Transmission of vector quantised 2-d DCT coefficients can then reduce the overall data rate to a level comparable with that for LPC filter parameters, but without having to satisfy the assumptions of AR analysis.

EXPERIMENTS

Sentences taken from the TIMIT database were analysed to produce maximum likelihood spectrograms which were then encoded with an algorithm similar to that proposed by the Joint Photographic Image Expert Group (JPEG) for images [2]. The quantisation was performed to deliver varying degrees of compression. As an example, figures 1 and 2 show spectrograms of a typical sentence before and after 2-d DCT coding, with a compression of 91%.

Figure 1: Spectrum of a sentence taken from TIMIT.

Figure 2: Reconstructed speech spectrum, original shown in figure 1, after using 2-d DCT.

The opinion equivalent quality [3] of the reconstructed speech was assessed for each level of compression, and then, intelligibility was assessed using the diagnostic rhyme test (DRT) [4], with both sets of results being shown in table 1. The DRT tests were performed with a small ‘in-house’ database.

Effective Compression ratio	Intelligibility	Opinion Equivalent Quality
Using 2-d DCT		
94%	96.6%	20dB
92%	97.2%	23dB
91%	97.9%	24dB
90%	98.6%	25dB
87%	98.9%	30dB
83%	99.0%	31dB
Using 10^{th} order LPC		
89%	96.3%	19dB

Table 1: Measures of effective compression ratio, intelligibility, and opinion equivalent quality of the respective coding methods.

CONCLUSIONS

A new approach to speech coding has been proposed, which takes account of frame-to-frame interdependency by treating the sequence of frames as a two-dimensional pattern and encoding it in a similar fashion to an image. This allows non-parametric representations of speech to be encoded at data rates comparable with parametric methods such as LPC [5]. The main advantage of this approach is the improved intelligibility and perceived quality of the reconstructed speech, while maintaining the low data rate of parametric methods.

REFERENCES

[1] DELLER J. R. JR., PROAKIS J. G. & HANSEN J. H. L., *Discrete-Time Processing of Speech Signals*, Macmillan, 1993.

[2] WALLACE G. K., "The JPEG still picture compression standard", *Communications of the ACM*, vol. 34, no. 4, pp. 31-44, April 1991.

[3] "IEEE recommended practice for speech quality measurements", *IEEE Trans. on Audio and Electroacoustics*, pp. 227-246, September 1969.

[4] VOIERS W. D., "Diagnostic evaluation of speech intelligibility" in M. E. Hawley (ed.), *Speech Intelligibility and Speaker Recognition,* Stroudsburg, Pa., Dowden Hutchenson and Ross, pp. 374-387, 1977.

[5] TREMAIN T. E., "The Government standard linear predictive coding algorithm: LPC-10", *Speech Technology*, vol. 1, pp. 40-49, April, 1982.