Proceedings of the Institute of Acoustics

FEATURE EXTRACTION VIA THE BICEPSTRUM FOR RECOGNITION OF NOISY SPEECH

H. Fairhurst, C.C. Goodyear.

University of Liverpool, Dept. of Electrical Engineering and Electronics, Brownlow Hill, Liverpool. L69 3BX.

1. INTRODUCTION.

Higher-order statistical techniques are attracting considerable current interest as a means for robust estimation of the parameters of noisy signals. For example, third-order cumulants are insensitive to added white or coloured Gaussian noise [1, 2], and this insensitivity extends to other noise sources with symmetrical probability density functions. These techniques therefore offer promise for improving the performance of speech recognisers when presented with noisy speech.

A technique [2] that has been used previously is to obtain an overdetermined set of linear equations from the third-order cumulant plane and apply a least squares technique to derive the AR parameters at a chosen order. The power cepstral coefficients are then found from these and used to form the recognition feature vectors. The method requires choices to be made regarding the order of the AR model and the initial cumulant equations, and the results may be sensitive to such choices.

This paper describes the use of an alternative numerical technique that works for MA, AR or ARMA signals, does not require knowledge of the model or its order or the selection of starting equations. The method employs a two-dimensional deconvolutional technique [3] to compute the complex cepstrum from the third-order cumulants of the speech signal; from this an estimate of the power spectrum can be made.

2. POWER SPECTRUM ESTIMATION VIA THE COMPLEX BICEPSTRUM.

2.1 Complex Bicepstral Method for Power Spectrum Estimation.

Let a stationary non-Gaussian i.i.d random sequence, $w(k)$, excite an ARMA system to produce the output sequence $x(k)$. The third-order cumulants of $x(k)$, which for a zero-mean signal are the same as the third-order moments, are given by:

$$R_x(m,n) = \overline{x(k)x(k+m)x(k+n)} \qquad (1)$$

FEATURE EXTRACTION VIA THE BICEPSTRUM

The two-dimensional Z-transform of $R_x(m,n)$ may be denoted by $B_x(z_1,z_2)$, which when evaluated on the unit surface gives the bispectrum. Pan and Nikias [3] introduced the complex bicepstrum, $c_x(m,n)$, of x(k) as:

$$c_x(m,n) = Z_2^{-1}\{\log[B_x(z_1,z_2)]\} \tag{2}$$

where Z_2^{-1} denotes the inverse two-dimensional Z-transform.

It may be shown [3] that :

$$c_x(m,n) = \begin{cases} \log|\beta| & m=0, n=0 \\ -\frac{1}{n}A^{(n)} & m=0, n>0 \\ -\frac{1}{m}A^{(m)} & n=0, m>0 \\ \frac{1}{m}B^{(-m)} & n=0, m<0 \\ \frac{1}{n}B^{(-n)} & m=0, n<0 \\ -\frac{1}{n}B^{(n)} & m=n>0 \\ \frac{1}{n}A^{(-n)} & m=n<0 \\ 0 & \textit{otherwise} \end{cases} \tag{3}$$

where β is the skewness of the input sequence w(k) and $A^{(n)}$ and $B^{(n)}$ are the complex cepstral coefficients of the impulse response of the ARMA system. Thus the complex bicepstrum is determined by the complex cepstrum along three straight lines and is zero elsewhere (see figure 1).

It is also shown [3] that a direct relationship between the complex bicepstrum sequence $c_x(m,n)$ and its third-order cumulant sequence, $R_x(m,n)$, exists and is formulated by a linear convolution equation:

$$R_x(m,n) * [mc_x(m,n)] = mR_x(m,n) \tag{4}$$

where * denotes the convolution operator.

Since the equation holds over the unit surface [4] a solution for $c_x(m,n)$ can be obtained by applying 2-dimensional discrete Fourier transforms (2D-DFT) :

$$c_x(m,n) = \frac{1}{m}F_2^{-1}\left\{\frac{F_2[m.R_x(m,n)]}{F_2[R_x(m,n)]}\right\} \tag{5}$$

where $F_2[.]$ and $F_2^{-1}[.]$ respectively denote 2D-DFT and inverse 2D-DFT.

FEATURE EXTRACTION VIA THE BICEPSTRUM

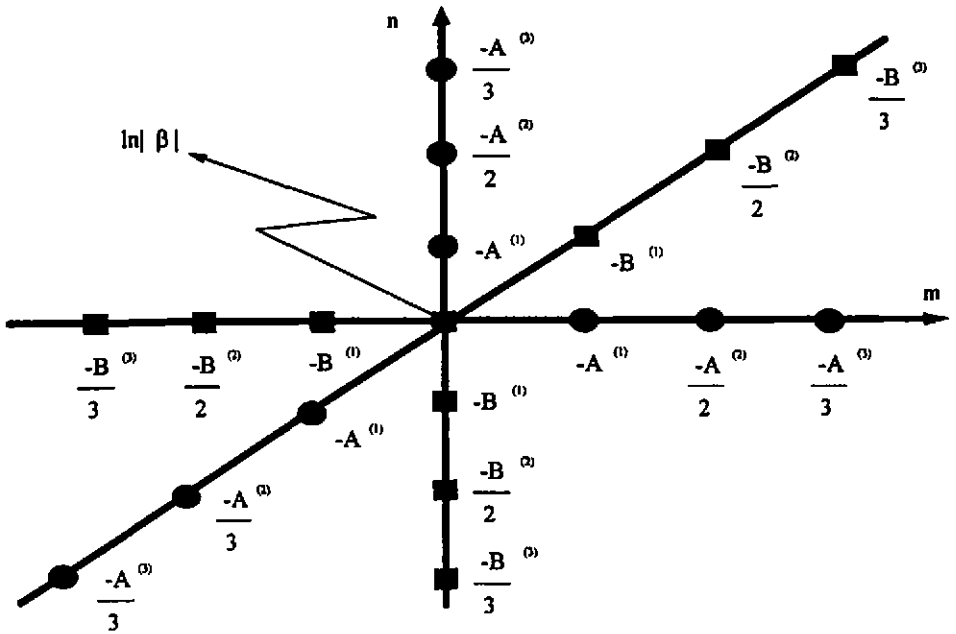

Figure 1. The complex bicepstrum of a non-minimum-phase deterministic signal (from [3])

It is seen in figure.1 that the complex cepstral A and B coefficients can be recovered from the complex bicepstrum by the following relations:

$$\begin{aligned} c_x(m,n) &= -\frac{1}{m}A^{(m)} \quad n=0;\, 0<m<M \\ &= \frac{1}{m}B^{(-m)} \quad n=0;\, -M<m<0 \end{aligned} \tag{6}$$

where M is the chosen window length of the estimated complex cepstrum sequence. We note that the method requires no phase unwrapping to compute complex cepstrum.

The power cepstrum sequence, $p(m)$, is related to its complex cepstrum sequence by the following relationship [4]:

$$p(m) = p(-m) = -\frac{A^{(m)} + B^{(m)}}{m} \qquad 0 < m \leq M \qquad (7)$$

The log power spectrum, $P'(\omega)$, can then be recovered by applying a FFT to the power cepstrum sequence, after zero-padding, to give an up-sampled version of the cepstrally smoothed log power spectrum :

$$P'(\omega) = F_1\{p(m)\} \qquad (8)$$

where $F_1[.]$ denotes DFT.

2.2 Comparison of Second-order (Fourier transform) and Third-order Power Spectrum Estimation Methods.

To demonstrate the robustness of the third-order power spectral estimation technique, a comparison of the estimated power spectra from the second and third-order methods is shown in figure 2. The estimate of the power spectrum by the third-order method was computed from a 25ms segment of the test signal. A two-dimensional Parzen window was applied to the estimated cumulant plane to select ±64 cumulant lags. Equations (5) and (6) were then used to generate 32 complex cepstral coefficients which by equation (7) gives 32 power cepstral coefficients. An estimate of the log power spectrum was then obtained by equation (8) from an appropriately zero padded version of the 31 power cepstral coefficients. To facilitate the second and third-order spectral comparison, the conventional (i.e. second-order) Fourier transform estimate of the power spectrum was cepstrally smoothed. This smoothing was achieved by selecting the same number of second-order power cepstral coefficients that were used in the third-order power spectral estimate (i.e. 32 coefficients).

The test signal used for the power spectrum comparison was a 25ms segment of a synthetic vowel 'AH' that had been generated from a 10th order LP model excited by an impulse train. Figure 2(a) compares the estimated spectra in the noiseless case. It can be seen from this graph that the third-order method gives a power spectral estimate that is a close match to the conventional or second-order estimate. Figures 2(b) and 2(c), respectively, show the estimated power spectra for a noisy vowel 'AH' by the second and third-order methods. The noisy vowel has been obtained by corrupting the synthetic vowel, by the addition of white Gaussian noise, to a signal-to-noise ratio (SNR) of 15dB. If these graphs are compared it can be seen, quite clearly, that the third-order estimate has been only marginally affected by the additive noise, whereas, the second-order estimate has lost its third formant completely and its second formant is starting to be swamped by the noise.

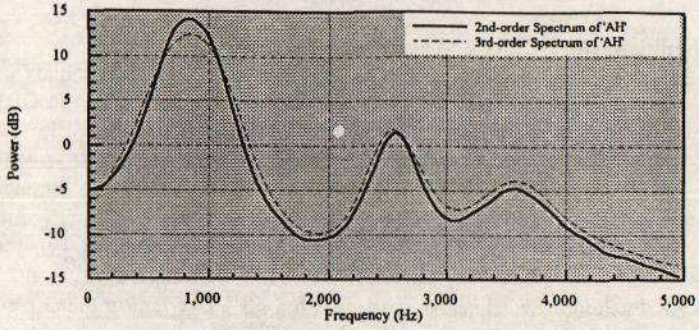

Figure 2(a) Power spectra for clean vowel 'AH'

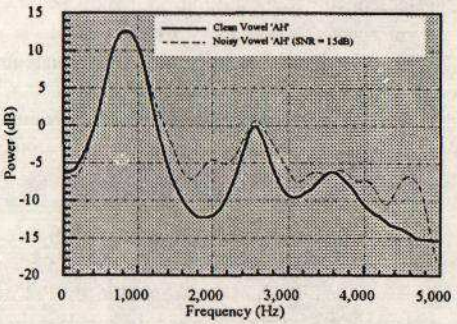

Figure 2(b) Second-order power spectrum estimates of vowel 'AH'.

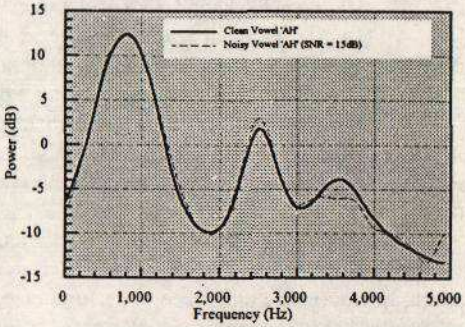

Figure 2(c) Third-order power spectrum estimates of vowel 'AH'.

Figure 2. Comparison of second and third-order power spectral estimates of synthetic vowel 'AH'

3. SPEECH RECOGNITION EXPERIMENTS.

We have performed two recognition experiments on speech with additive noise. The first experiment investigated the robustness of the third-order technique for speech that has been corrupted by the addition of white Gaussian noise. The second experiment investigated the robustness of the third-order technique for speech that has been corrupted by car noise.

As stated in the introduction to this paper, third-order cumulants are insensitive to Gaussian processes [1,2], whether white or coloured. For this reason the recognition experiments that this paper addresses are for vowel recognition, since vowel sounds fit better into the class of a non-Gaussian signal than do fricative sounds.

3.1 Vowel Recognition with Additive White Gaussian Noise.

In this experiment the recognition task was to classify steady-state vowel sounds into ten vowel classes. Ten vowels were selected (AH, IY, ER, AE, OO, EH, II, UH, OH and OR), to generate five hundred CVC words uttered by a single speaker. These utterances were lowpass filtered to 4.4kHz and sampled at 10kHz. A 25ms segment of the steady-state part of the vowel of each CVC utterance was manually located and extracted, giving a total of fifty utterances for each vowel.

The feature vector type chosen to represent each speech segment was composed of mel-scale frequency cepstral coefficients (MFCC's). Each feature vector was generated from the estimated power spectrum by placing a triangular filter bank, of order twenty, over the frequency range of 0 - 5kHz. Each filter was linearly spaced with a 50% overlap on the corresponding mel-scale frequency. A discrete cosine transform (DCT) was then applied the log aggregate energy of each filter to generate eight MFCC's.

A maximum likelihood (ML) classifier was selected for the recognition task. The ML classifier classifies the input feature vector, x, into vowel class i if $p(x|\ i) > p(x|\ j)$ for all $j \neq i$, where $p(x|\ i)$ is the probability density function (PDF). The PDF assumed for this particular classifier was multivariate Gaussian assuming a diagonal covariance matrix.

To perform the recognition experiment the data base was divided into twenty five training tokens and twenty five test tokens for each vowel. White Gaussian noise produced from a pseudo-random number generator was added to the clean test vowels at particular SNR (SNR range 5 - 50dB).

The results for the recognition experiment are given in figure 3. It can be seen from figure 3 that the second-order method gives better performance than the third-order method on clean speech, however, this difference in performance is somewhat smaller than that reported by Paliwal and Sondhi [2] for their method. We note, however, that figure 3 relates to a single speaker data-base whereas the results in [2] are multi-speaker. In noisy conditions it can be seen that the third-order method gives better robustness than the second-order method. At 15dB SNR the recognition accuracy has fallen to approximately 49% for the second-order method, whereas the third-order method is approximately 74% accurate.

3.2 Vowel Recognition with Additive Car Noise.

In a second set of experiments noise was taken from recordings made in a moving car (file IDO5 on the ETSI database), lowpass filtered to 4.4kHz. Figure 4(a) shows the power spectrum of a 25mS segment of the car noise and figure 4(b) shows the PDF for the noise.

The experimental procedure for the recognition tests was identical to that in the white noise tests. Again, clean test vowels were corrupted by the additive noise to obtain chosen SNR values. The SNR calculations were made over the signals full bandwith of 4.4kHz and without applying any form of weighting filter. The results for the recognition experiment are shown in figure 5. Again, at high SNR (20dB) the second-order method gives a marginal improvement in recognition accuracy compared to the third-order method, but at low SNR the converse is true. At -5dB SNR the recognition accuracy has fallen to approximately 49% for the second-order method, whereas the third-order method is approximately 62% accurate. The relative improvement in recognition

accuracy of the second and third-order methods in the presence of additive car noise compared to that in additive white Gaussian noise is probably due to the way in which SNR is calculated.

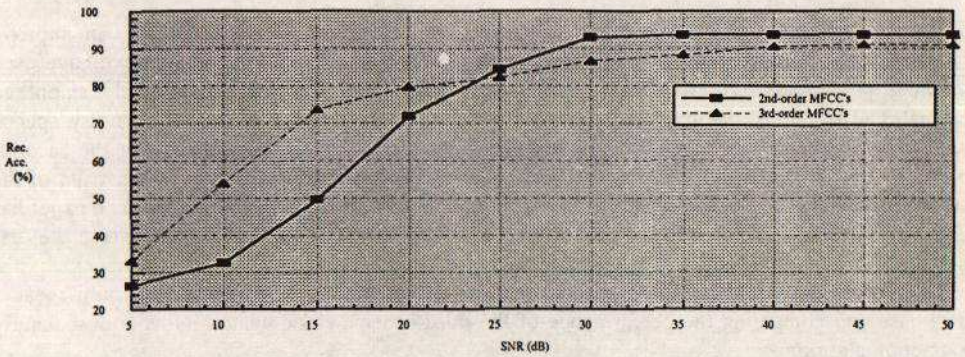

Figure 3. Recognition accuracy versus signal-to-noise ratio for vowel recognition in white gaussian noise.

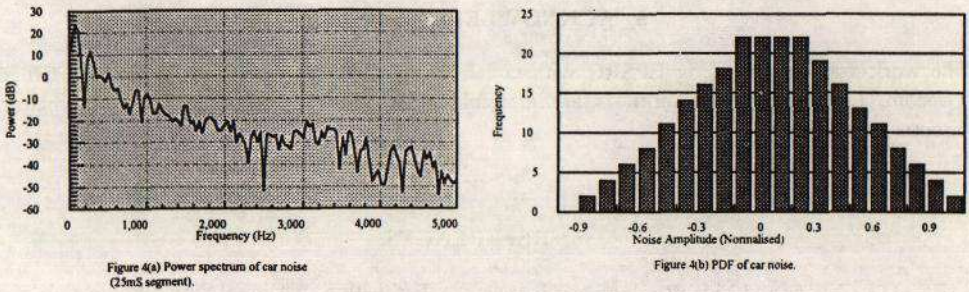

Figure 4(a) Power spectrum of car noise (25mS segment).

Figure 4(b) PDF of car noise.

Figure 4. Power spectrum and Probability Density Function (PDF) of car noise.

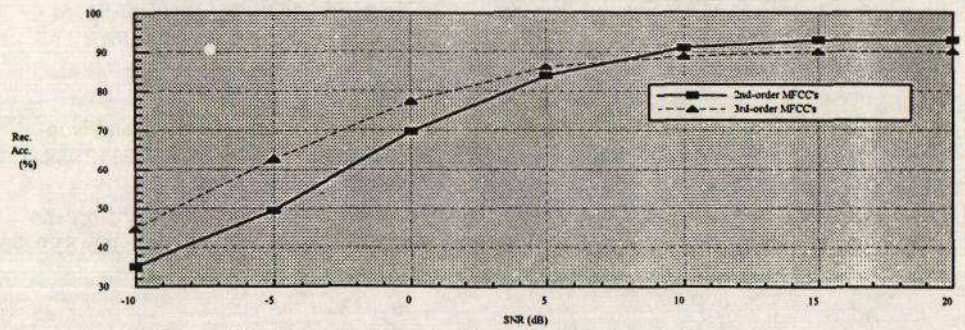

Figure 5. Recognition accuracy versus signal-to-noise ratio for vowel recognition in car noise.

4. CONCLUSIONS.

This paper has demonstrated that feature extraction via the complex bicepstrum can improve recogniser performance when presented with noisy speech. The third-order method we have used provides power spectral estimates which match closely the conventional cepstrally smoothed estimates. The third-order method offers useful recognition performance gains on noisy speech without seriously degrading the performance on clean speech, when compared with the second-order method. It is also notable that the third-order method recovers the complex cepstrum of the speech signal, and thus preserves the non-minimum phase information of the speech. The paper has also demonstrated that the third-order method gives robustness to a 'real' type of noise, i.e. car noise, and without the need to first estimate the noise characteristics.

Future work will include extending the vowel recognition experiments to a multi-speaker data-base and comparing the performance of the third-order method against other robust feature extraction algorithms.

5. ACKNOWLEDGEMENTS.

The work was supported by EPSRC with collaboration under the CASE scheme with British Telecom. Laboratories. The authors are grateful to Dr. Andrew Lowry of BTL for helpful discussions.

6. REFERENCES.

[1] C.L. Nikias and M.R. Raghuveer, "Bispectrum Estimation: A Digital Signal Processing Framework", Proc. IEEE, vol. 75, pp.869 - 891, 1987.

[2] K.K. Paliwal and M.M. Sondhi, "Recognition of Noisy Speech using Cumulant-Based Linear Predicition Analysis", Proc. IEEE Int. Conf. Acoust. Speech & Signal Proc. ICASSP '91, ppI - 429 - 432, 1991.

[3] R. Pan and C.L. Nikias, "The Complex Cepstrum of Higher-order Cumulants and Non-minimum Phase System Identification", IEEE trans. on ASSP, vol 36, No. 2, Feb 1988.

[4] C. L. Nikias and L. Fang, "Bicepstrum Computation based on Second and Third-order Statistics with Applications", IEEE Int. Conf. Acoust. Speech & Signal Proc. ICASSP '90, pp2381 - 2385, 1990.