A FORMANT VOCODER BASED ON MIXTURES OF GAUSSIANS

Parham Zolfaghari *Tony Robinson*

Cambridge University Engineering Department, Trumpington Street, Cambridge CB2 1PZ, UK.
Tel: [+44] 1223 332754 Fax: [+44] 1223 332662
email : psz1000,ajr@eng.cam.ac.uk

ABSTRACT

This paper describes a new low bit-rate formant vocoder. The formant parameters are represented by Gaussian mixture distributions, which are estimated from the discrete Fourier transform (DFT) magnitude spectrum of the speech signal [12]. A voiced/unvoiced classification mechanism has been developed based on the harmonic nature of each formant in the DFT spectrum modulated by the Gaussian Mixture distribution. Using a magnitude-only sinusoidal synthesiser [8], intelligible synthetic speech has been obtained. Vector quantisation [3] of the vocal tract parameters enables this formant vocoder to operate at a bit-rate of 1248 bps.

1. INTRODUCTION

Most methods for analysing and synthesising speech are based on a parametric description of the short-time spectrum or an equivalent representation of the speech signal. In this paper a new parametrisation method is used in a speech analysis/synthesis process in order to develop a formant vocoder. In the past, formant based speech coders have been shown to work at very low bit rates [4, 11], as the spectral content of speech can be represented by only three or four formants.

Formants characterise many speech sounds and represent the resonances of the vocal tract. Identification of formants depends on several factors, for example, the method used to obtain the smoothed magnitude function, the relative amplitudes and frequencies of the resonances, the method used to pick the peaks and the parameters used for processing the speech signal. Model based techniques such as Linear Prediction analysis are used in order to extract formants but problems such as spurious peaks exist.

The magnitude of the discrete Fourier transform (DFT) directly contains the formant information and can serve as a basis for formant analysis of speech. A formant extraction technique has been developed whereby the short-time magnitude spectrum is modelled by probability density functions represented by mixtures of Gaussians [12]. The EM (Expectation Maximisation) algorithm [2] is used to perform the parameter estimation process.

This mixtures of Gaussians based technique has been integrated into a low bit-rate vocoder system, as shown in Figure 1, whereby the probability density function (pdf) parameters are encoded and decoded. Sinusoidal coders have produced good quality synthesised speech and as a closer match to the analysis procedure, the magnitude-only sinusoidal synthesis model allows the reconstruction of speech from Gaussian mixture parameters.

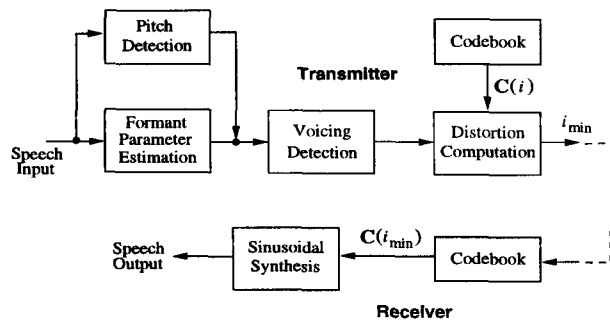


Figure 1. Diagram of Formant Vocoder Structure.

The rest of the paper is formalised as follows. First the mixtures of Gaussians based formant analysis technique is reviewed, followed by a description of the developed voiced/unvoiced classification and pitch detection mechanism. The vector quantisation (VQ) process and codebook generation methods are then summarised. Finally the synthesis system is described and results from analysis and synthesis of speech using this vocoder are presented.

2. FORMANT ANALYSIS USING GAUSSIAN MIXTURES

A new formant analysis technique using Gaussian mixtures to depict the short time spectral structure of speech has been developed. The EM algorithm for finding the maximum likelihood of a mixture model is used to perform the parameter estimation process. The means, variances and mixture weights of the probability density functions represent the formant frequencies, bandwidths and amplitudes respectively. Figure 2 shows an estimated mixture distribution of four Gaussians superimposed over the DFT magnitude spectrum that is obtained by analysis of a short segment of speech.

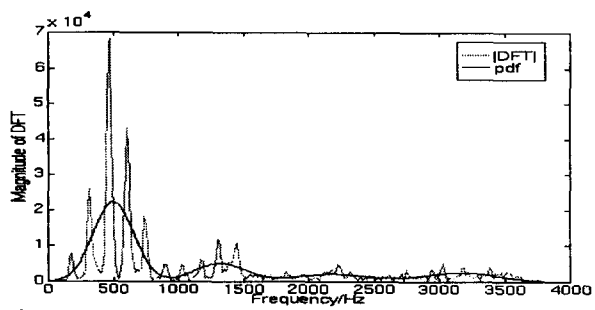
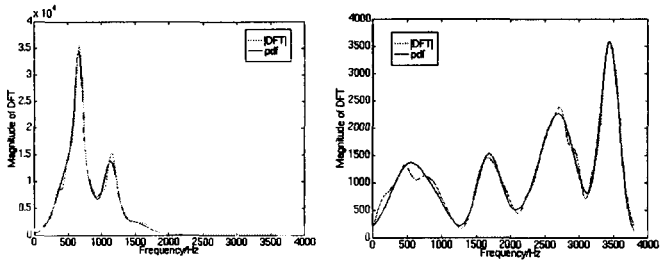


Figure 2. A mixture of Gaussians fit (pdf) to a DFT magnitude spectrum.



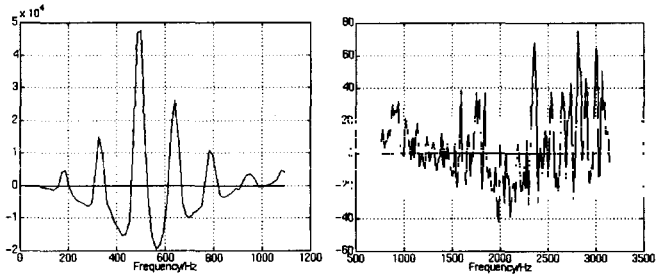
(a) (b)

Figure 3. Plots of Gaussian mixtures (pdf) superimposed on two cepstrally smoothed spectrums, where a) is for a voiced segment and (b) is an unvoiced section of speech.

As can be seen, the spectral shape has been well represented and the formants within the frequency range have been picked by each of the Gaussians in the mixture. In order to attempt to better fit the mixture distributions to the DFT magnitude spectra the effect of varying several factors was investigated based on simple models of speech perception. Cepstral smoothing [1] of the the magnitude spectra has produced the best formant parameters when compared to the original speech signal. An example of this is shown in Figure 3. A spectrogram of the sentence "*we were away a year ago*" is shown in Figure 5, which also illustrates a spectrogram representation of the Gaussian mixtures per frame. The fits were obtained after the application of cepstral smoothing to magnitude spectra. The latter clearly shows the formant tracks obtained using this analysis technique.

3. PITCH & VOICING DETERMINATION

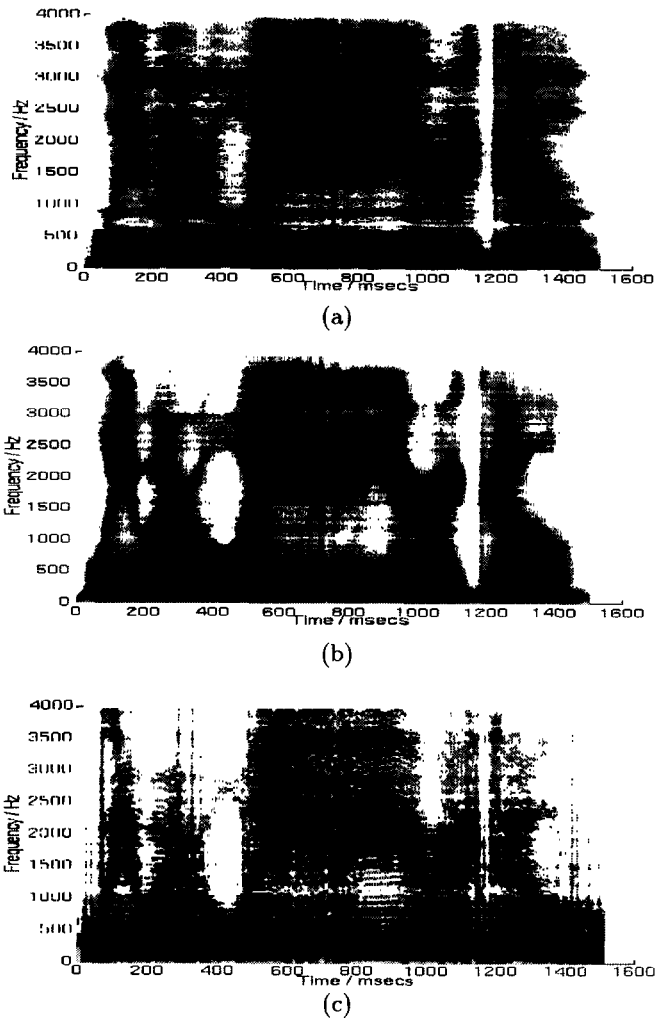
In order to complete a speech analysis/synthesis system, pitch estimation and voicing classification are required. In the systems developed, pitch detection was performed using the cepstrum method as described in [9, 10]. This method requires the computation of the cepstrum for a short segment of speech which is then searched for the peak cepstral value and its location. The pitch period is the location of the peak between fixed thresholds.



(a) (b)

Figure 4. Plots of (a) voiced formant and (b) unvoiced formant for zero-crossing calculations.

An algorithm for voiced/unvoiced classification per formant has been developed. This makes use of the harmonic nature of formants to decide on voicing. In Figure 2 note that the formants are modulated by the Gaussian distribution. Figure 4 shows the difference between the spectrum of the formants and the Gaussian distribution for a voiced and an unvoiced segment of speech. These show that a correlation exists between the number of zero-crossings per



(a)

(b)

(c)

Figure 5. Figure (a) shows the spectrogram of the original speech, (b) represents the estimated probability density functions spectrogram using mixtures of Gaussians and (c) shows formant synthesised speech for the same utterance using the Klatt synthesiser.

formant and the classification of voicing. As the pitch is known, the actual number of zero-crossings within the formant can be compared with the number calculated, and this can serve as the basis for a voiced/unvoiced decision per formant. This method has been found to be successful and easy to implement.

4. VECTOR QUANTISATION

Standard vector quantisation was utilised in this vocoder. Three codebooks were trained each of dimensionality four (one dimension per Gaussian). Separate codebooks were trained for the mean, standard deviation and mixture weight parameter vectors. The means and mixture weights were transformed to the log domain before quantisation, and similarly the standard deviation was represented as a fraction of the mean. Table 1 shows these spectral parameter vector conversions for construction of the training data.

The VQ codebooks were built from training data comprising of frames of formant analysed natural speech of 58

different speakers lasting approximately 43 minutes. Each codebook was trained using the LBG algorithm [6].

Acoustic analysis was done at a fixed frame rate of 32 msec resulting in a bit rate of 1248 bps.

For simplicity each frame was coded independently of adjacent frames, and relaxing this constraint is expected to lead to lower bit-rates.

Parameter Type	Representation	Bit Allocation
4 x Mean	Logarithm	10
4 x Std Deviation	Fraction of Mean	7
4 x Weight	Logarithm	10
4 x Voicing	-	4
Pitch	Reciprical	8
Total bits per Segment		39

Table 1. Conversion of Parameters and Bit Allocation

5. THE SYNTHESIS SYSTEMS

In the initial phase of the vocoder design the Klatt synthesiser [5] was used for formant synthesis. In order to drive this synthesiser, the pdf parameters (means, variances and mixture weights) were converted to formant parameters. Despite the shape of a normal distribution not being directly related to the second order filters commonly used in formant synthesisers, results from the Klatt synthesiser have yielded intelligible synthetic speech, although a large number of parameters are required. Figure 5(c) shows an example utterance synthesised using the Klatt synthesiser.

Sinusoidal modelling enables the reconstruction of the time waveform from the parametrised speech as a closer match to the analysis procedure described.

5.1. The Sinusoidal Synthesis Model

McAulay described a sinusoidal model for the speech waveform [7], for which the phase is defined as the integral of the instantaneous frequencies of the component sine waves. From classical speech perception the assumption can be made that the ear is sensitive principally to the short-time spectral magnitude and not the phase, provided that phase continuity is maintained.

The speech waveform can be modelled as a sum of sine waves. If $s(n)$ represents the sampled speech waveform then

$$s(n) = \sum_i A_i(n) \sin[\phi_i(n)] \qquad (1)$$

where $A_i(n)$ are the amplitudes and $\phi_i(n)$ is the time-varying phase of the i'th partial. The phase is taken to be the integral of the instantaneous frequency $f_i(n)$ and thus can be shown to satisfy the recursion

$$\phi_i(n) = \phi_i(n-1) + 2\pi n f_i(n)/f_s \qquad (2)$$

where f_s is the sampling frequency. As the pitch is known the partials are harmonically related by

$$f_i(n) = i * f_0(n) \qquad (3)$$

where $f_0(n)$ is the fundamental frequency at time n. As a consequence of the definition of phase in terms of the instantaneous frequency, waveform continuity is obtained. Each frequency and amplitude of the constituent sinusoids was linearly interpolated on a sample by sample basis. This

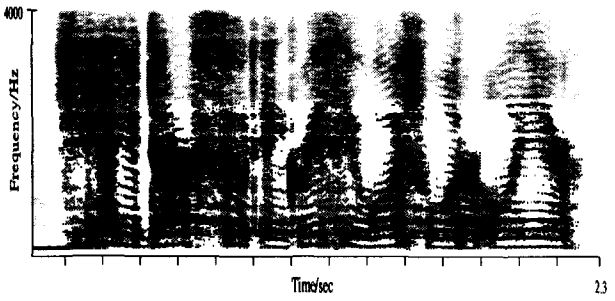


Figure 6. Spectrogram of the original utterance *"She had your dark suit in greasy wash water all year"*.

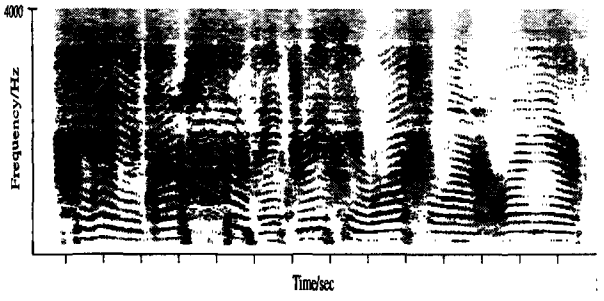


Figure 7. Spectrogram derived by the magnitude-only synthesis technique for the same utterance as that of Figure 6.

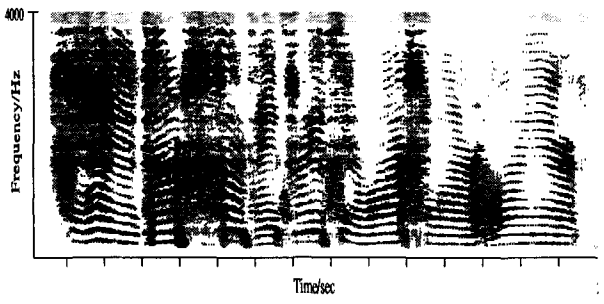


Figure 8. Spectrogram derived by quantising the parameters.

results in considerably better partial tracks as shown in Figure 7 with original speech shown in Figure 6 (audio file is provided). Note that the reconstructed phase function is not the same as the original speech waveform but this is perceptually toned down provided that the magnitude spectrum has been successfully reconstructed.

The simplest waveform reconstruction method, the overlap-add method, was also tested. Using a triangular window enables the amplitudes and the frequencies to be linearly interpolated across frame boundaries resulted in intelligible synthetic speech but discontinuities occurred on some sinusoids across frame boundaries.

6. SYSTEM EVALUATION

Figure 8 shows a spectrogram of utterance *"She had your dark suit in greasy wash water all year"* after quantisation of the spectral parameters. An example of female speech is shown in Figures 9 and 10. Note that none of the spectral parameters from these sentences were in the codebook training data.

In informal intelligibility tests, 5 sentences after quantisation of the parameters were presented to listeners, all of whom found all this speech intelligible, although some requested a second hearing as the sentences had specialist vocabulary out of context.

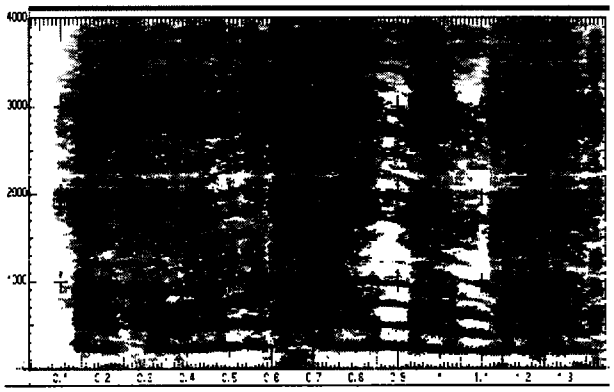


Figure 9. Spectrogram of Original Speech.

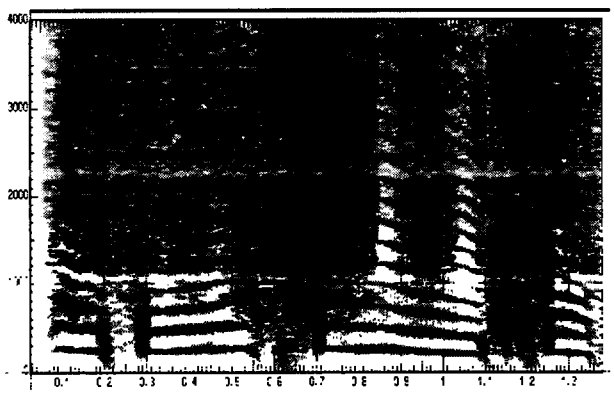


Figure 10. Spectrogram from magnitude-only synthesis technique after quantisation of the parameters.

7. CONCLUSIONS & FURTHER WORK

The results presented here have further developed the effectiveness of mixtures of Gaussians in speech synthesis and analysis. A formant vocoder has been developed using a simple vector quantiser to encode the spectral parameters. An operating bit-rate of 1248 bps was obtained.

With comparison to other methods of analysis, the Gaussian means as well as the variances between frame boundaries slowly vary in time and this fact will be employed in designing a better coding mechanism such as segmental coding, for operation at even lower bit-rates.

In the sinusoidal synthesis model the phase was defined in terms of the instantaneous frequency, hence reconstruction depends only on the amplitudes and frequencies of the component tones. This model still requires modifications in order to better model the formant structure represented by the mixtures of Gaussians. At the moment there is a suspected mismatch between the gain of the spectrum and the Gaussian mixture distribution. This results in less resonant formant frequencies reducing the quality of synthesised speech.

An advantage to this technique in comparison to linear prediction analysis is that the number of parameters is independent of the sample rate. In linear prediction the parameters tend to increase with higher sampling rates. This would allow better modelling of speech with lower number of parameters.

ACKNOWLEDGEMENTS

P.S. Zolfaghari is funded by an EPSRC studentship.

REFERENCES

[1] J.R. Deller, J.G. Proakis, and J.H.L. Hansen. *Discrete-time Processing of Speech Signals.* Macmillan Publishing Company, 1993.

[2] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society Series B*, 39:1–38, 1977.

[3] R. M. Gray. Vector quantisation. *IEEE ASSP Magazine*, 1(2):4–29, April 1984.

[4] C. Jaskie and B. Fette. A survey of low bit rate vocoders. *Proceedings of Voice Systems Worldwide*, February 1992.

[5] D.H. Klatt. Software for a cascade/parallel formant synthesiser. *Journal of the Acoustical Society of America*, 67(3):971–995, March 1980.

[6] Y. Linde, A. Buzo, and R. M. Gray. An algorithm for vector quantiser design. *IEEE Transactions on Communications*, 28:84–95, Jan 1980.

[7] R.J. McAulay and T.F. Quatieri. Magnitude-only reconstruction using a sinusoidal speech model. In *Proc. Int. Conf. Acoust., Speech and Signal Processing*, page 27.6.1, 1984.

[8] S.H. Newab, T.F. Quatieri, and J.S. Lim. Signal reconstruction from short-time Fourier transform magnitude. *IEEE Trans. Acoust., Speech and Signal Processing*, ASSP-31(4):986–998, August 1983.

[9] L.R. Rabiner, M.J. Cheng, A.E. Rosenberg, and C.A. McGonegal. A comparative performance study of several pitch detection algorithm. *IEEE Trans. Acoust., Speech and Signal Processing*, ASSP-24(5):399–418, October 1976.

[10] R.W. Schafer and L.R. Rabiner. System for automatic formant analysis of voiced speech. *Journal of Acoustical Society of America*, 49:1867–1873, 1970.

[11] Nigel Sedgwick. Emulation of a formant vocoder at 600 and 800 bps. In *EUROSPEECH 93*, pages 523–526, 1993.

[12] P. Zolfaghari and A.J. Robinson. Formant Analysis using Mixtures of Gaussians. In *International Conference on Spoken Language Processing*, Oct 1996.