FORMANT ANALYSIS USING MIXTURES OF GAUSSIANS

Parham Zolfaghari *Tony Robinson*

Cambridge University Engineering Department,
Trumpington Street, Cambridge CB2 1PZ, UK.

ABSTRACT

This paper describes a new formant analysis technique whereby the formant parameters are represented in the form of Gaussian mixture distributions. These are estimated from the Discrete Fourier Transform (DFT) magnitude spectrum of the speech signal. The parameters obtained are the means, variances and the masses of the density functions, which are used to calculate centre frequencies, bandwidths and amplitudes of formants within the spectrum. In order to better fit the mixture distributions various modifications to the DFT magnitude spectrum, based on simple models of perception, were investigated. These include reduction of dynamic range, cepstral smoothing, use of the Mel scale and pre-emphasis of speech. Results are presented for these as well as formant tracks from analysing speech using the final formant analysis system.

1. INTRODUCTION

The spectral envelopes of many speech sounds are characterised by several prominent maxima. These represent the resonances of the vocal tract and are called *formants*. Formants are of interest to us for a number of reasons: they represent the most immediate source of articulatory information and the source most familiar to us by virtue of our use of formant information in speech perception. Hence, formant estimators, which are used in applications such as speech coding, either implicitly or explicitly, examine the spectral envelope.

The problem of automatic formant analysis has received considerable attention during the past two decades, and a variety of approaches have been explored [6]. Schafer and Rabiner [7] presented the first detailed approach for automatically estimating formant structure from voiced speech using cepstral analysis. Markel [4] has presented a simplified procedure for estimating the formant frequencies using linear prediction techniques. Most of the above techniques are widely used in speech analysis and perform a deconvolution to separate the impulse response and the glottal driving function.

The Fourier transform, in both analog and discrete-time forms, has been the basis for many important developments in speech analysis and synthesis. Clearly, the DFT can serve as a basis for formant analysis of speech, since it directly contains the formant information in its magnitude spectrum. Once the DFT of the speech signal has been calculated then other techniques can be applied to obtain the constituent components of speech.

This paper introduces a new formant analysis technique whereby Gaussian mixture distributions are fitted to discrete Fourier Transform magnitude spectra. The EM (Expectation Maximisation) algorithm is used to perform the parameter estimation process. The rest of the paper is organised as follows. The EM algorithm is briefly described in section 2. It is then implemented for the problem of formant analysis in the same section and some results are presented. In section 3 various spectral modifications for improving formant estimation is discussed which is then followed by results and conclusions.

2. MIXTURE OF GAUSSIANS FOR FORMANT ESTIMATION

Firstly the EM algorithm is reviewed here followed by a discussion of its application to formant parameter estimation using Gaussian mixtures.

2.1. The EM Algorithm

This section outlines the basic learning algorithm for finding the maximum likelihood of a mixture model [2], [5]. The EM algorithm is used to estimate Gaussian mixture distributions from histograms of DFT magnitude data. Viewing the spectrum as a probability density function the E-step computes the expected complete data likelihood. For this model this step requires the computation of the likelihood and the posterior probability for each bin in the histogram resulting from each mixture. The M-step re-estimates the means and variances of the Gaussians using the data set after accumulating sufficient statistics in order to maximise the likelihood. Initially the total area under the histogram is calculated; at this point the means, variances and mixture weights are initialised. The means are initialised uniformly over the interval. The variances are made significant with respect to the interval and the number of Gaussians in the mixture, and finally the mixture weights are set equal values.

The EM algorithm can be used with most distribution types but Gaussians are generally used in standard applications of the algorithm. The equation for a Gaussian mixture distribution is as follows:

$$b_j(\underline{x}) = \sum_{m=1}^{M} c_{jm} \frac{1}{\sqrt{2\pi}\underline{\sigma}_{jm}} e^{-\frac{1}{2}\left[\frac{\underline{x}-\underline{\mu}_{jm}}{\underline{\sigma}_{jm}}\right]^2} \quad (1)$$

where c_{jm} is the mixture weight for each of the M mixtures, with

mean $\underline{\mu}_{jm}$ and standard deviation $\underline{v}_{jm}$. Note also that

$$\sum_{m=1}^{M} c_{jm} = 1 \; ; \; c_{jm} \geq 0 \qquad (2)$$

In the following sections the implementation of the formant analyser using the above algorithm is outlined with some results.

2.2. Formant Analysis Technique

The speech signal may be considered to be stationary over short periods of time. For short time spectral analysis to be carried out, it is first necessary to window the signal so as to reduce the edge effects at the beginning and the end of the frame. A Hamming window of length 16ms was used with a frame advance of 8ms. Following this, the Fourier transform magnitude spectrum is obtained for each such frame.

Fitting Gaussian mixtures to the DFT magnitude spectra enables the estimation of the spectral features, and this was achieved by use of the EM algorithm described above. The means, variances and mixture weights of the density functions may then be used to calculate the formant frequencies, bandwidths and amplitudes. Figure 1 shows an estimated mixture distribution of four Gaussians superimposed over the DFT magnitude spectrum that is obtained by analysis of one frame of speech. As can be seen, the four formants within the frequency range have been picked by each of the Gaussians in the mixture. However, various problems were encountered while

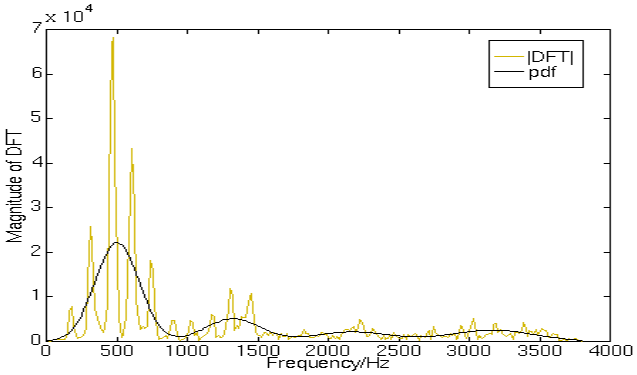
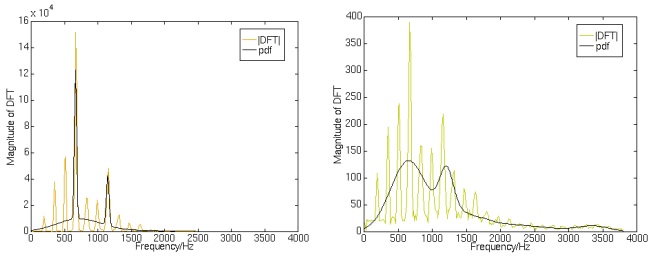


Figure 1: A Mixture of Gaussians fit to a DFT magnitude spectrum.

fitting Gaussians to certain formants such as the case where two formants are close together in frequency. In order to better fit the mixture distributions to the magnitude spectra the variation of several factors was investigated based on simple models of perception. The following section reviews these modifications to the spectra and presents the results obtained.

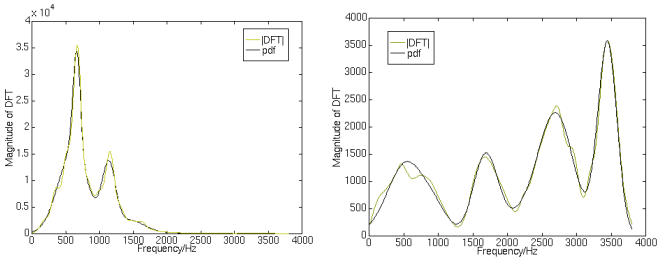
3. MODIFICATIONS TO THE SPECTRUM

Although in some voiced frames the formant frequencies were obtained from the corresponding Gaussian distribution mean (Figure 1), in other cases their associated bandwidths were rather small as the distribution would just pick the harmonic at the formant frequency and neglect the adjacent harmonics associated with that formant resulting in a very small variance (Figure 2(a)). Two methods



(a) (b)

Figure 2: Plots of Gaussian mixtures superimposed on a) DFT magnitude spectrum and b) square root of the DFT magnitude spectrum.



(a) (b)

Figure 3: Plots of Gaussian mixtures superimposed on two cepstrally smoothed spectrums, where a) is a smoothed version of Figure 2(a) and (b) is an unvoiced section of speech.

were investigated to resolve this problem which occurs frequently with Gaussian mixture fits to magnitude spectra and they are:

- The dynamic range of the power spectra was reduced using a non-linear amplitude scale such as the square root or the cube root of the spectra. Firstly the cube root of the spectra were modelled but the results obtained were not as good as expected since the distinction in spectral features in the DFT was considerably reduced in some cases. Thus the square root of the spectra was used resulting in a better representation of spectral features. Figure 2 shows a comparison of Gaussian mixture representations of the original linear dynamic range and after the square root of the same DFT magnitude spectrum has been taken. Although the resulting formant frequencies are the same, the bandwidth of the formants are better modelled using this modification of the spectrum.
- The application of cepstral smoothing [1], which is the process of removing the high-quefrency effects of the excitation from the spectrum. Figure 3(a) shows the cepstral smoothed spectrum of Figure 2(a) with Gaussian mixture fit superimposed. Figure 3(b) shows that unvoiced sections of speech can be also modelled well using Gaussian mixtures.

After experiments on various utterances using the above modifications of the spectra it was deduced that the cepstral smoothing technique gave considerably better fits and smoother formant trajectories, specifically for the second and third formants.

From studies of the human ear we know that the human auditory system does not perceive pitch in a linear manner. It has been shown that it is favourable to have increased frequency resolution at lower

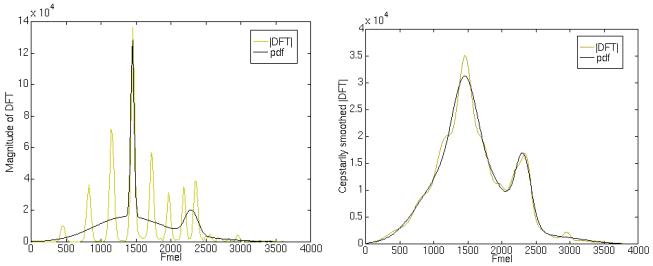
frequencies, for example the Mel scale is a unit of measure of perceived pitch or frequency of a tone. Also voiced speech spectra normally have a drop-off of about 6-dB/Octave, which results in a high spectral dynamic range. In effect, the speech spectra are tilted into a slightly low-pass form. In order to reduce this effect, the speech signal is often pre-emphasised to increase the relative energy of the high frequency spectrum. Mixture densities were fitted to the above modifications of the magnitude spectra, and in the following sections results obtained from each is given.

3.1. Formant Analysis Using the MEL-Scale

Figure 4(a) shows a frame with its magnitude spectrum warped using the Mel scale described by:

$$F_{mel} = 1125 \log[1 + \frac{F_{Hz}}{625}]. \quad (3)$$

This technique does introduce some problems however. As can be seen in Figure 4(a) one Gaussian has been allocated to the lower frequency harmonics of the first formant. This occurred frequently using this analysis technique, and thus the cepstral smoothing technique described above was used in order to overcome this problem as shown in Figure 4(b). Although the Mel scale allowed better resolving of the lower frequency formants, it was found that the higher frequency formants were not picked as well as with no warping of the frequency scale, as at higher frequencies the formants were compressed closer together. Thus, this technique was not used in the final system.



(a) (b)

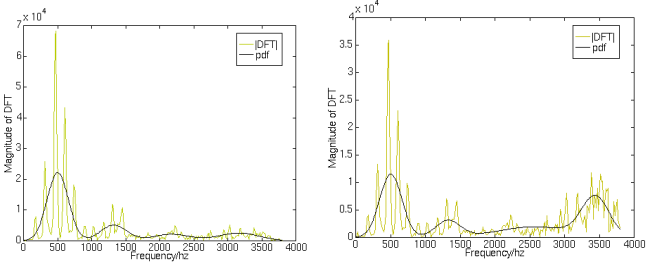
Figure 4: Plot (a) shows Gaussian mixtures superimposed on DFT magnitude spectrum after warping using the Mel Scale, and b) shows the same analysis frame after cepstral smoothing.

3.2. Formant Analysis Using Pre-emphasis

Finally, in order to increase the relative energy of the higher frequencies, the input speech was pre-emphasised. The pre-emphasis filter has the following form:

$$P(z) = 1 - \mu z^{-1}. \quad (4)$$

where μ was chosen to be 0.97. Figure 5(b) shows the result of the Gaussian mixture fits after pre-emphasis of the speech signal. This figure illustrates that the formant at 3500Hz is picked after pre-emphasis, but the formant at 2220Hz has been neglected. If Figure 5(a) is looked at closely it can be seen that both these formants have been allocated a Gaussian before pre-emphasis, and thus by using pre-emphasis the formant estimates are worsened. This is the problem with pre-emphasising the signal where in some frames Gaussians are wrongly allocated to higher frequencies. This technique was not used in the final formant analysis system.



(a) (b)

Figure 5: Plots of Gaussian mixtures superimposed on a) DFT magnitude spectrum with no pre-emphasis, and b) magnitude spectrum with pre-emphasis.

4. CONVERSION OF GAUSSIAN PARAMETERS

In this section an outline of conversion of means, variances and mixture weights to formant parameters is given. The formant frequencies are obtained from the means of each Gaussian and the amplitudes are the probability density functions at these centre frequencies. The 3-dB bandwidth for each formant is calculated from the corresponding Gaussian in the probability density function. Referring to equation 1 the mean μ_{jm} and variance σ_{jm} in each Gaussian in the mixture is known, then by using the 3-dB log ratio, the bandwidth BW of the Gaussian distribution can be calculated as:

$$BW = 2\sigma_{jm}\sqrt{\ln[2]} \quad (5)$$

5. RESULTS

So far cepstral smoothing of the DFT magnitude spectrum have produced the best formant parameters in comparison to the original signal. It was also deduced that higher number of Gaussians in the mixture resulted in better fits to the spectrum. A spectrogram of the sentence "*we were away a year ago*" is shown in Figure 6, 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 the technique described. The bandwidths are slightly larger than the original which is due to the application of cepstral smoothing. Figure 8 is a similar representation of the sentence "*the rain in Spain falls mainly on the plain*". Although the bandwidths are larger the formant tracks are well represented.

In order to compare the Gaussian mixture formant tracks to other formant tracking methods, an LPC based formant tracker, namely ESPS waves+, was used. In this tracker the formant frequencies are selected from candidates proposed by solving for the roots of a 14th order linear predictor polynomial computed periodically from the speech waveform. Figure 7 shows the resulting formant tracks from both the LPC based system and the Gaussian mixture system. Again this diagram shows agreeable formant trajectories. Note that the LPC based system uses dynamic programming in order to optimise formant trajectory estimates by imposing frequency continuity

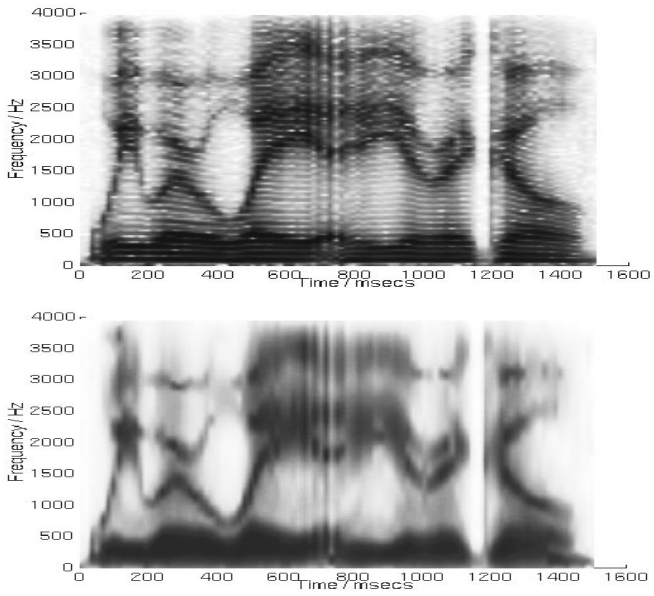


Figure 6: Top diagram shows the spectrogram of the original speech and the bottom diagram represents above view of estimated probability density function spectrogram.

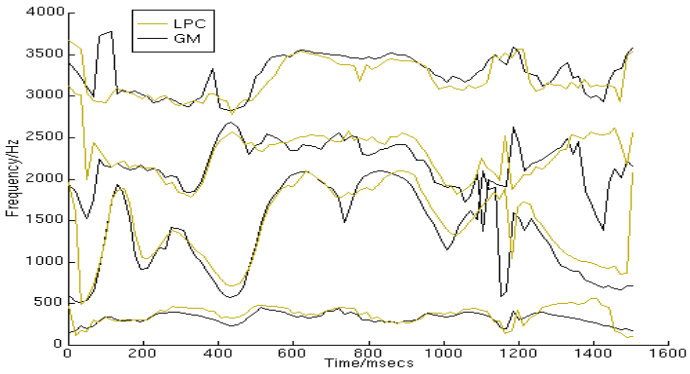


Figure 7: Formant tracks from linear prediction and the Mixtures of Gaussians techniques.

constraints, and as yet there are no frame to frame constraints in this Gaussian mixture system.

Although the shape of a normal distribution is not directly related to second order filters commonly used in formant synthesisers, results from the Klatt synthesiser [3] have yielded intelligible synthetic speech.

6. CONCLUSIONS & FUTURE WORK

The results presented here have demonstrated the effectiveness of Gaussian mixtures in estimating formant parameters. Further refinements to this system require the devise of a formant trajectory smoothing algorithm in order to exploit known smoothness constraints. An application for this technique is low bit rate speech coding. In Figure 5(a) it can be observed that in the lower frequency Gaussians a set of pitch pulses have been convoluted by a window

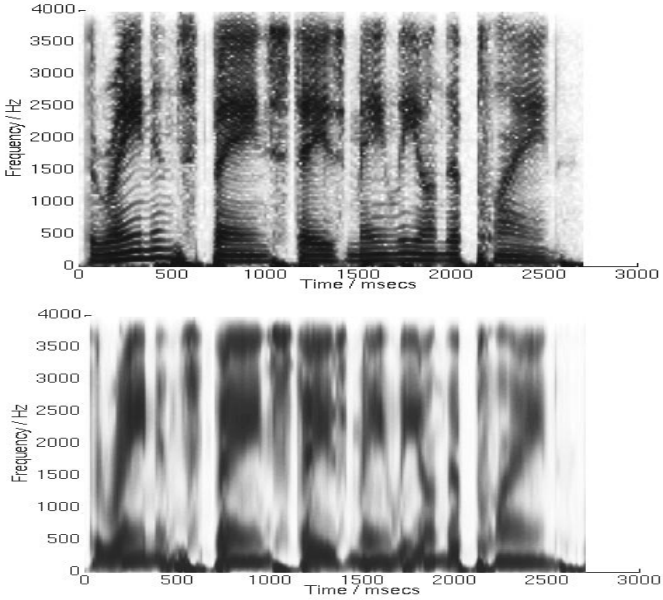


Figure 8: Top diagram shows Spectrogram of original speech and the bottom diagram represents above view of estimated probability density function spectrogram.

function and that at the higher frequencies it is largely unvoiced. This information can be built into a coder. Also, other methods of reconstructing the time waveform from the parametrised speech that more closely match the analysis procedure are to be investigated.

Acknowledgements

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

7. 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. D.H. Klatt. Software for a cascade/parallel formant synthesiser. *Journal of the Acoustical Society of America*, 67(3):971–995, March 1980.
4. J.D Markel. Basic formant and Fo parameter extraction from a digital inverse filter formulation. *IEEE Transactions*, AU-21:69–79, 1973.
5. R.A. Redner and H.F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, 1984.
6. R. Schafer and J. Markel. *Speech Analysis*. New York: IEEE Press, 1979.
7. 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.