SPEECH CODING USING MIXTURE OF GAUSSIANS POLYNOMIAL MODEL

Parham Zolfaghari[*] *Tony Robinson*[†]

[*]CREST/ATR Human Information Processing Research Labs, Kyoto 619-02, Japan
email : zparham@hip.atr.co.jp
[†]Cambridge University Engineering Department,Cambridge CB2 1PZ, UK
email : ajr@eng.cam.ac.uk

10.21437/Eurospeech.1999-340

ABSTRACT

We have investigated a novel method of spectral estimation based on mixture of Gaussians in a sinusoidal analysis and synthesis framework. After quantisation of this parametric scheme a fixed frame-rate coder operating at a bit-rate of around 2.4 kbits/s has been developed. This paper describes an extension to this spectral model based on constraining the parameters of the mixture of Gaussians to be on a polynomial trajectory over a segment of speech data. This is referred to as the mixture of Gaussians polynomial model (MGPM). In order to realise a segmental coder, dynamic programming over the utterance is performed. The segmental representation of the spectra results in a log-likelihood score over a segment which is used as the cost function in the dynamic programming algorithm. Speech coding components such as pitch, voicing and gain are described segmentally. A number of segmental coders are presented with bit-rates in the range of 350 to 650 bits/s. These coders offer good and intelligible coded speech evaluated using DRT scoring at these bit-rates.

1. INTRODUCTION

A segmental framework employs the inter-frame or time dependence of the spectral representation. This dependence is inherent in various segments of speech, such as sustained vowels, as the speech spectral envelope is a slow time-varying process and spectra of adjacent frames are highly correlated. Various forms of segmentation models have been applied to speech coding and speech recognition. In speech coding Roucos *et al* [11] describe a very low bit-rate segment vocoder operating at 150 bits/s for a single speaker. This low rate is achieved by vector quantisation (VQ) of all the LPC spectra in a segment as a single unit. The Kang-Coulter 600 bits/s vocoder [6] also uses LPC methods followed by formant tracking to produce good quality speech with a reported DRT score of 79.9. These low bit-rates can also be achieved by a recognition-based approach where recognition units are coded. Holmes [5] has described a method which uses an underlying linear-trajectory formant model for both recognition and synthesis.

The contribution of this work is to model the envelope of the short-term power spectral density as a mixture of Gaussians [13]. In this framework a Gaussian roughly corresponds to a formant with the Gaussian mean corresponding to the formant frequency and the variance corresponding to the bandwidth. This model was integrated in a sinusoidal model based speech coding scheme [14]. An advantage of this framework is that a speech segment may be modelled using a polynomial trajectory for the Gaussian means and variances. We have previously reported on a segmental coder using a linear polynomial trajectory for the Gaussian mixtures operating between 600-800 bits/s [15]. We extend this model to an R'th order polynomial to represent both means and variances of the Gaussians. In the speech recognition area, similar models have also been implemented for MFCC trajectories in a HMM-based system [4].

2. SEGMENTAL CODER STRUCTURE

The block structure of the coders described in this paper is as shown in Figure 1. A sinusoidal model framework based on the ideas of McAulay and Quatieri [8] is used. In this model, the speech signal is represented by a harmonic set of partials with varying amplitudes and frequencies. In accordance with our desire to build a very low bit-rate coder we restrict the sine waves to be harmonically related. The inverse FFT method of re-synthesis [3] is used and the phase of each harmonic is chosen at reconstruction time to minimise the mismatch with the previous frame.

The Spectral Envelope Estimation Vocoder (SEEVOC) envelope, devised by Paul [9] uses a robust peak detection algorithm to yield a smooth envelope as the underlying spectral representation. In order to operate in the low bit-rate region, the SEEVOC envelope needs to be efficiently coded. We aid the mixture of Gaussians polynomial model to represent this spectra over a segment. Polynomial least squares

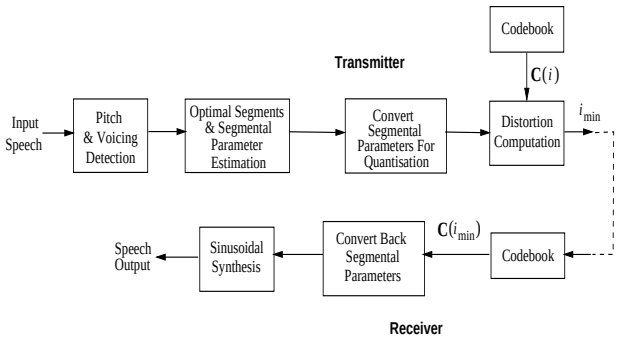
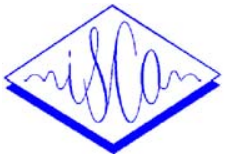


Figure 1. Block diagram of MGPM vocoder structure.

fits of pitch and gain over the segment, and a segmental representation of voicing is used to complete the parametric segmental system. The following sections describe these segmetnal representations in more detail.

2.1. Mixture of Gaussians Polynomial Model

Mixture models are extensively used in applications where data can be viewed as arising from several populations mixed in varying proportions. This maximum likelihood estimation technique can be applied to learn parameters of the mixture pdf. The Expectation Maximisation (EM) algorithm [2] is a broadly applicable algorithm used to maximise the log-likelihood from incomplete data, by iteratively maximising the expectation of log-likelihood from complete data. The SEEVOC magnitude spectrum, once normalised is considered as a probability distribution $P_t(k)$, where k are the bin numbers ($k = \{1, \cdots, N\}$) and t is the frame index. Hence, $P_t(k)$ is simply a normalised spectral density at time frame t. For this model we use the EM algorithm as an optimisation tool for estimating mixture density parameters, viewing $P_t(k)$ as a histogram.



6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

2.1.1. Derivation of Model Parameter Estimation Formulae

Let a data sample k be the observed incomplete data and $(P_t(k), y_k)$ be the complete data, where y_k is an unobservable integer between 1 and s. This term y_k then indicates the number of density components $f(k|y_k, \phi_{y_k})$ and mixing parameters ω_{y_k} of the mixture pdf, where the parameter set is given by $\mathbf{\Phi} = \{\phi_1, \cdots, \phi_s\}$.

The aim is simply to fit the histogram representation of the spectrum where the means of the densities are FFT bin numbers and are restricted to a polynomial over a segment of spectra. This results in mixture densities which represent a smoothed spectral shape for all t. We assume that a parametric family of mixture probability density functions $f(k,t|\mathbf{\Phi})$ is given and that $\bar{\mathbf{\Phi}}$ represents the parameter values to be estimated. The log-likelihood of the data set can be formulated as follows

$$\mathcal{L}(\mathbf{\Phi}) = \log \left[\prod_{t=1}^{T} \prod_{k=1}^{N} f(k,t|\mathbf{\Phi})^{P_t(k)} \right] \tag{1}$$

The posterior probability is represented by

$$P(y_k|k,t,\mathbf{\Phi}) = \frac{\omega_{y_k} f(k,t|y_k,\phi_{y_k})}{\sum_{y_k} \omega_{y_k} f(k,t|y_k,\phi_{y_k})}. \tag{2}$$

The Q-function can be represented as:

$$\begin{aligned} Q(\mathbf{\Phi}, \bar{\mathbf{\Phi}}) &= \sum_i \left\{ \sum_{t=1}^{T} \sum_{k=1}^{N} \gamma_i(k,t) \right\} \log \bar{\omega}_i \\ &+ \sum_i \left\{ \sum_{t=1}^{T} \sum_{k=1}^{N} \gamma_i(k,t) \log f(k,t|i,\bar{\phi}_i) \right\} \end{aligned} \tag{3}$$

where $\gamma_i(k,t) = P_t(k)P(i|k,t,\phi_i)$. Since there is a summation over all y_k $(1 \leq y_k \leq s)$, y_k is independent of k, and therefore, can be denoted by i.

Maximisation of the Q-function is obtained by maximising each term of this function for each mixture i with respect to the mixture weight ω_i and the mixture parameters ϕ_i. This is the EM update equation for mixture density equations which leads us onto the definition of the density to be used and maximisation of the associated parameters.

2.1.2. Maximisation

The mixture density used in this research is the Gaussian distribution and hence, before differentiating the above function, the functional terms within the equation are defined as follows:

$$f(k,t|i,\bar{\mathbf{\Phi}}) = \mathcal{N}(k, \bar{\mathbf{b}}_i, \bar{\sigma}_i^2) = \frac{1}{\sqrt{2\pi\bar{\sigma}_i^2}} \exp\left\{ -\frac{(k - \mathbf{z}_t\bar{\mathbf{b}}_i)^2}{2\bar{\sigma}_i^2} \right\} \tag{4}$$

where

$$\bar{\mathbf{b}}_i = [b_{i0}, b_{i1}, \ldots, b_{iR}]^T \tag{5}$$

$$\mathbf{z}_t = \left[1, t, t^2, \ldots, t^R\right]. \tag{6}$$

Thus, the mean of the Gaussian distribution described by equation 4 is represented by an R'th order polynomial. After differentiating equation 1 with respect to the mean trajectory parameters b_{ir} of the i'th mixture, and equating to zero, by substitution and noting that $\bar{\sigma}_i$ is independent of time we obtain

$$\sum_{t=1}^{T} \sum_{k=1}^{N} \gamma_i(k,t) \cdot kt^r = \sum_{t=1}^{T} \sum_{k=1}^{N} \gamma_i(k,t) \cdot t^r \sum_{u=0}^{R} \bar{b}_{iu} \cdot t^u \tag{7}$$

$$r = 0, \ldots, R$$

which results in a set of linear simultaneous equations which are solved for the mean trajectory parameters b_{ir}.

In order to estimate the variance of each mixture $\bar{\sigma}_i^2$, equation 1 is differentiated with respect to this variance and solved for $\bar{\sigma}_i^2$ resulting in the maximum likelihood estimate given by:

$$\bar{\sigma}_i^2 = \frac{\sum_{t=1}^{T} \sum_{k=1}^{N} \gamma_i(k,t)(k - \mathbf{z}_t\bar{\mathbf{b}}_i)^2}{\sum_{t=1}^{T} \sum_{k=1}^{N} \gamma_i(k,t)}. \tag{8}$$

Finally, the mixture weights $\bar{\omega}_i$ can be estimated using equation 1 by application of a Lagrange multiplier [1] resulting in the following

$$\bar{\omega}_i = \frac{\sum_{t=1}^{T} \sum_{k=1}^{N} \gamma_i(k,t)}{\sum_{t=1}^{T} \sum_{k=1}^{N} \sum_{j=1}^{M} \gamma_j(k,t)}. \tag{9}$$

This completes the EM based maximisation of the parameter space of the mixture of Gaussians with polynomial mean trajectories.

2.1.3. Variance Trajectory Model

It is also possible to obtain a more precise model of the spectral dynamics within the segment by representing the variance of the i'th mixture by an R'th order polynomial where the likelihood of the data given the parameters is given by

$$\mathcal{N}(k, \bar{\mathbf{b}}_i, \bar{\mathbf{c}}_i) = \frac{1}{\sqrt{2\pi\mathbf{z}_t\bar{\mathbf{c}}_i}} \exp\left\{ -\frac{(k - \mathbf{z}_t\bar{\mathbf{b}}_i)^2}{2\mathbf{z}_t\bar{\mathbf{c}}_i} \right\}$$

where

$$\bar{\mathbf{c}}_i = [c_{i0}, c_{i1}, \ldots, c_{iR}]^T$$

$$\mathbf{z}_t = \left[1, t, t^2, \ldots, t^R\right].$$

After differentiation of equation 1 with respect to $\bar{c}_{ir}$ and setting to zero, a non-linear equation in $\bar{c}_{in}$ is obtained. In order to make this a linear function, an approximation is used where the $\bar{c}_{in}$ in the denominator of the maximised log-likelihood is replaced by the current value, c_{in} [4, 12].

An illustration of a mixture of Gaussians polynomial fit over a short segment of speech is given in Figure 2.

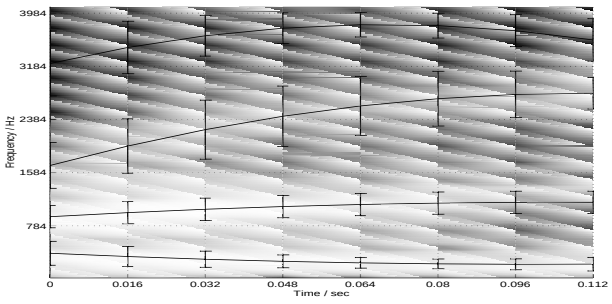


Figure 2. Spectrogram representation of MGPM tracking using both mean and variances, over a short segment of speech with 4 Gaussians.

2.2. Initialisation

The mixture of Gaussian means are uniformly initialised over the interval. The variances are made significant with respect to the interval and the number of Gaussians in the mixture. The mixture weights are set equal values. In addition to this, for the higher order polynomial components (i.e. not the 0'th component), an initial value of 0.01 is set. The convergence point is on average at around 9 iterations.

2.3. Segmental Pitch & Gain

The gain requires a considerable number of bits to code per segment if coded on a single frame basis. More efficient compression can be achieved by using a least means square polynomial fit over a segment. The pitch determination algorithm is based on the autocorrelation method. Representation of pitch over a segment is also based on a polynomial fit, using polynomials of order two for pitch and order three for gain. Figure 3 demonstrates polynomial fits over an utterance after segmentation. Note that both these components were dynamically transformed using a logarithmic scale.

2.4. Segmental Voicing

Voicing is detected using the autocorrelation function in combination with the low-band and high-band energy. When the segment boundaries have been obtained, the following segmental vector representation of voicing within the segment is used. Initially, a maximum number of transitions of voicing is set. At the start of the segment the voicing in the transmitter and the receiver is set to 1 (voiced). Using the voicing decisions within the length of the segment, the voicing transitions are represented as a fraction of this segment length. As an example, if the maximum number of transition is five and at the start, the segment is voiced, a vector can be constructed as follows:

$$v = [1, 0, 0, 0, 0] \tag{10}$$

If there are voicing transitions at frames four (to unvoiced) and nine (to voiced), and the maximum segment length is twelve, then the vector would be represented as,

$$v = [1, 0.3333, 0.75, 0, 0] \tag{11}$$

which results in an efficient vector representation of the voicing within the segment boundaries.

3. OPTIMAL SEGMENTATION

A simple method of segmenting speech has been devised based on dynamic programming. The cost function is the log-likelihood of the segment described by equation 1. A transition penalty is assigned in order to penalise short segment lengths. Various experiments were carried out in order to find the optimum transition penalty for obtaining the best segment boundaries from a speech coding point of view. This view places less constraint on the accuracy of the segment boundaries as no classification step is required.

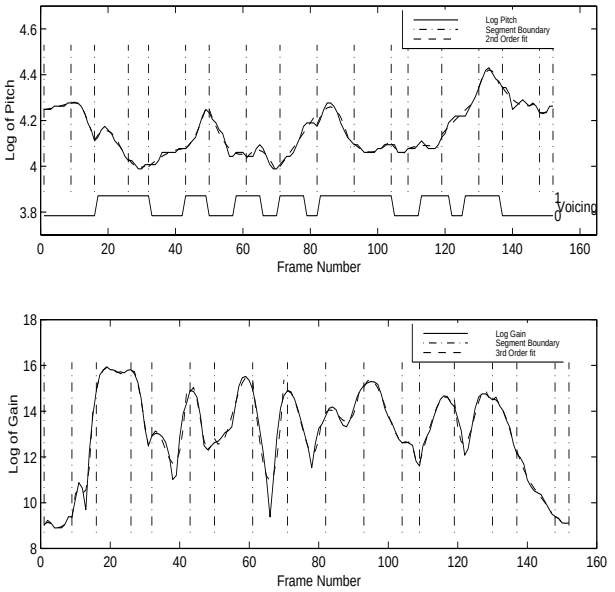


Figure 3. Gain and Pitch polynomial fits.

4. CODER IMPLEMENTATION

A number of coders were devised based on the above parameterisation techniques. Speech is sampled at 8 kHz and a fixed frame-rate of 16 ms is used. First, the speech signal is broken down into frames, with a frame-step of 128 samples. A hamming window of length 512 is used. This is then transformed using a 512 point FFT, yielding a 256 point spectrum.

Two methods of vector representations for coding the MGPM parameters are possible. As an example, consider the quantisation of the mean trajectory parameters (Equation 5). The case where there are six Gaussians within the mixture with polynomials of order $R = 2$, results in a m by n matrix where $m = 6$ and $n = 3$. Vectors for quantisation can be constructed based on either a $m\times$row or $n\times$column representation of this matrix. By using a column vector representation, a considerable reduction in the number of vectors is obtained and as a consequence the number of codebooks are also reduced. The MGPM codebooks were trained based on this column vector representation. The vector b_{k0} was warped based on the Mel-scale in quantisation, where k is the mixture component index. Also the standard deviation vector c_{k0} was represented as a fraction of b_{k0}.

The VQ codebooks were constructed from training data obtained from WSJCAM0 corpora [10]. Natural speech from 50 different speakers lasting approximately 39 minutes in total was used. Each codebook was trained using the LBG algorithm [7].

4.1. Bit Allocation of MGPM coders

After segmentation, the average segment length of the words in DRT score was found to be 9.4 frames using a constant transition penalty. This results in an average frame-rate of 6.65 frames/sec. Table 1 below summarise the bit allocation for coders to obtain average bit-rates of 650, 550, 450, and 350 bits/s. The coder resulted in a DRT score of 88.8 using no quantisation of the parameters. The coder operating at 650 bits/s is the only coder that is using the variance trajectory model. High number of bits is allocated on the first column vectors of the mixture of Gaussians polynomials which include the b_{k0} and c_{k0} vectors. A total of 99 bits/segment was allocated.

We found that in order to be able to operate at 550 bits/s and achieve high intelligibility the variance trajectories needed to be discarded. The number of bits required is too high to justify using this model for these lower bit-rates. Thus we assume that the formant bandwidths within the segment are a constant.

In order to operate below 450 bits/s, rather than reduce the number of bits for each parameter type, better compression of speech was achieved by increasing the average segment length to 10.5 frames. This results in an average frame-rate of 5.95 frames/sec over the DRT words which was achieved by decreasing the transition penalty from 0.01 to 0.006.

Figure 4 illustrates spectrograms of "There are always problems with new plans he said" using various levels of compression of this segmental coding scheme [Sound files attached].

5. PERFORMANCE OF MGPM CODERS

Subjective evaluations have been based on twelve listeners DRT scores, with every person presented with a different 96 stimuli words. Table 2 illustrates the scores obtained. It can be seen that as compared to the standard coders the intelligibility of the 350 bits/s coder is at the scale of the 2.4 kbits/s LPC-10e and the 650 bits/s is marginally higher than that of CELP-4.8.

6. CONCLUSIONS

The subject of this paper has been the investigation and formulation of a particular branch of digital speech cod-

Operating Bit-rate		650 bits/s	550 bits/s	450 bits/s	350 bits/s
Parameter Type	**Representation**	**Bits/Segment**	**Bits/Segment**	**Bits/Segment**	**Bits/Segment**
b_{k0}	Mel-scale	12	12	12	9
b_{k1}	-	12	10	8	7
b_{k2}	-	7	9	6	5
c_{k0}	Fraction of b_{k0}	12	12	10	8
c_{k1}	-	8	-	-	-
c_{k2}	-	7	-	-	-
Mixutre Weights	-	10	10	10	8
Segmental Gain	-	12	12	12	9
Segmental Pitch	-	9	9	9	7
Segmental Voicing	-	6	6	6	5
Segmental Duration	-	4	4	4	4
Total bits per Segment		99	84	77	62

Table 1. Representation of parameters and bit allocation per segment.

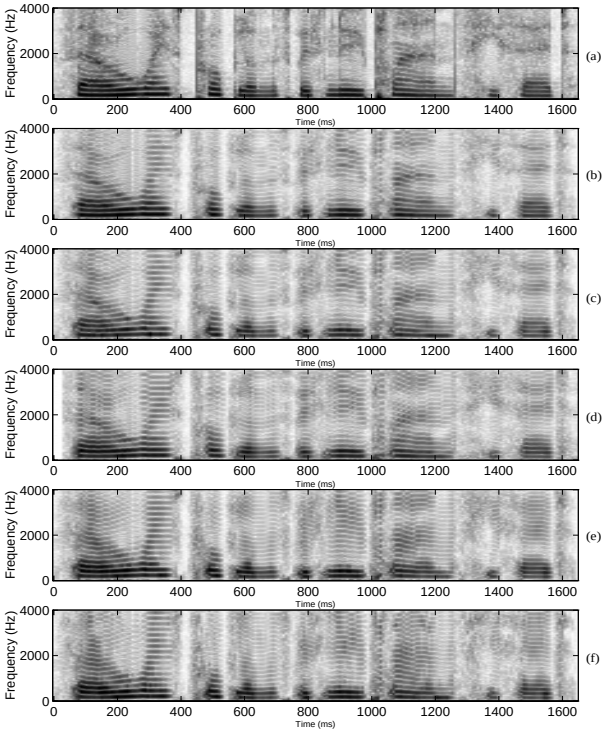


Figure 4. The above wide-band spectrograms represent (a) original 8 kHz input speech, (b) GM coded speech at 2450 bits/s, and MGPM coded speech at (c) 650 bits/s, (d) 550 bits/s, (e) 450 bits/s and (f) 350 bits/s.

Coder	Bit-rate	DRT score
LPC-10e	2400 bits/s	68.0
CELP	4800 bits/s	83.8
Fixed-Rate GM coder	2400 bits/s	88.1
Segmental GM coder	650 bits/s	84.4
Segmental GM coder	550 bits/s	81.2
Segmental GM coder	450 bits/s	72.5
Segmental GM coder	350 bits/s	67.4

Table 2. DRT results for the segmental and fixed-rate Gaussian Mixture (GM) coders

ing, namely sinusoidal model based segmental speech coding. The schemes developed have been formulated for very low bit-rate applications. The work focused on the extension of the fixed frame-rate Mixture of Gaussian based coder to a segmental parametric coder using mixture of Gaussians polynomial models. In order to obtain a segmental coder, dynamic programming using the best fit of the MGPM over segments resulted in segment boundary estimation. All speech coding components were described segmentally and quantised, resulting in coders operating at bit-rates of 350-650 bits/s. Decisions were made on the allocation of bits using listening tests to judge the importance of the segmental parameters. The new coder was compared against LPC10e and CELP using DRT evaluation and found to provide the same DRT score at only 15% of the bit rate.

REFERENCES

[1] C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press Oxford, 1995.

[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] P. Depalle and X. Rodet. Synthèse additive par FFT inverse. *Rapport Interne IRCAM*, 1990.

[4] T. Fukada, Y. Sagisaka, and K. K. Paliwal. Model parameter estimation for mixture density polynomial segment models. In *Proc. IEEE ICASSP'97*, volume 2, pages 1403–1406, 1997.

[5] W. J. Holmes. Low bit-rate speech coding using a linear-trajectory formant representation for both recognition and synthesis. *Proc. Instititute of Acoustics*, 20(6):179–186, 1998.

[6] G.S. Kang and D.C. Coulter. *600-Bit-Per-Second Voice Digitizer (Linear Predictive Formant Vocoder)*. Naval Research Laboratory, November 1976.

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

[8] R.J. McAulay and T.F. Quatieri. Magnitude-only reconstruction using a sinusoidal speech model. In *Proc. IEEE ICASSP'84*, page 27.6.1, 1984.

[9] D.B. Paul. The spectral envelope estimation vocoder. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-29(4):786–794, August 1981.

[10] A. J. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals. WSJCAM0: A British English speech corpus for large vocbulary continuous speech recognition. In *Proc. IEEE ICASSP'95*, pages 81–84, 1995.

[11] S. Roucos, M. Schwartz, and J. Makhoul. A segment vocoder at 150 b/s. In *Proc. IEEE ICASSP'83*, pages 61–64, Boston, April 1983.

[12] P. Zolfaghari. *Sinusoidal Model Based Segmental Speech Coding*. PhD thesis, Cambridge University, 1998. Submitted.

[13] P. Zolfaghari and A.J. Robinson. Formant Analysis using Mixtures of Gaussians. In *Proc. ICSLP'96*, volume 2, pages 1229–1232, Oct 1996.

[14] P. Zolfaghari and A.J. Robinson. A formant vocoder based on Mixtures of Gaussians. In *Proc. of IEEE ICASSP'97*, volume II, pages 1575–1578, April 1997.

[15] P. Zolfaghari and A.J. Robinson. A segmental formant vocoder based on linearly varying Gaussians. In *Proc. EUROSPEECH'97*, volume 1, pages 425–428, Rhodes, Greece, September 1997.