A SEGMENTAL FORMANT VOCODER BASED ON LINEARLY VARYING MIXTURE OF GAUSSIANS

Parham Zolfaghari and 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 low bit-rate segmental formant vocoder. The formants are estimated using mixture of Gaussians whose means are constrained to vary linearly with time within a segment. A new method of smoothing the power spectrum has been used in order to improve modelling with mixtures of Gaussians. Pitch is estimated using the autocorrelation function, and voicing is detected using the autocorrelation function method and the energy in the spectrum. Optimal segment boundaries are obtained using a dynamic programming procedure based on the power normalised log-likelihood of the segment. Magnitude-only sinusoidal synthesis is then used to synthesise speech from the estimated spectrum. Using multiple codebooks an average bit-rate of 500 bps has been obtained.

1. INTRODUCTION

The magnitude of the short-time discrete Fourier transform directly contains the formant information and can serve as a basis for the formant analysis of speech. In earlier work, a formant extraction technique was developed whereby the cepstrally smoothed short-time magnitude spectrum is modelled by a probability density function (pdf) represented by a mixture of Gaussians [7]. This technique was integrated into a low bit-rate formant vocoder system [8], whereby the pdf parameters are encoded and decoded. The magnitude-only sinusoidal synthesis [5] model has been adapted for speech synthesis using these mixture of Gaussians parameters. In this formant vocoder, the speech waveform is divided into frames with a fixed frame size, and for each frame and each vector within the frame, a fixed coding structure is used regardless of the local phonetic content.

However, for some segments of speech, especially for sustained vowels, the speech spectral envelope is actually a slow time-varying process, and spectra of adjacent frames are highly correlated. This concept has been utilised to design a segmental formant vocoder which would enable variable bit-rate operation and consequent reduction in the operating bit-rate. A linear model has been chosen to represent this correlation between frames within a segment.

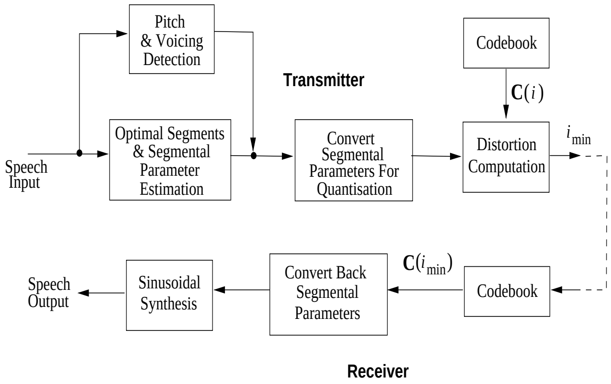
Pitch & Voicing Detection
Transmitter
Codebook
C(i)
Speech Input
Optimal Segments & Segmental Parameter Estimation
Convert Segmental Parameters For Quantisation
Distortion Computation
i_min
Speech Output
Sinusoidal Synthesis
Convert Back Segmental Parameters
C(i_min)
Codebook
Receiver

Figure 1: Diagram of Segmental Formant Vocoder Structure.

Various forms of segmentation models have been applied to speech coding and speech recognition. Roucos, Schwartz and Markhoul [6] describe a very low bit-rate segment vocoder operating at 150 bps 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 segmentation algorithm is a heuristic algorithm which uses a set of thresholds from two spectral derivatives to determine the spectral steady-state regions in the input speech. In Goldenthal's work [2], the temporal behaviour of phones is modelled by templates of dynamics of the acoustic attributes used to represent the signal. However, Goldenthal concentrates on modelling the trajectories of mel scale cepstral coefficients which have been shown [3] to oscillate through time even when the signal is changing slowly and smoothly. This paper presents a segmental trajectory model for formants within phonetic segments which are known to vary smoothly in time and are more directly related to the speech production mechanism.

This trajectory model is an extension of the mixture of Gaussians model for formant estimation. Within a segment, the mean of each Gaussian is constrained to vary linearly in time. Speech is segmented into a sequence of contiguous variable length segments using dynamic programming of the frame normalised log-likelihood.

This paper is structured as follows: After reviewing the segmental vocoder a detailed definition of the segmental model is presented. The optimal segment boundary estimation process is then described followed by a description of the synthesis procedure and VQ codebook generation. Finally the work is summarised and related future work proposed.

2. SEGMENTAL FORMANT VOCODER

The segmental formant vocoder of Figure 1 is a statistical method for the segmental modelling of speech in the power spectral domain. A new spectral smoothing technique is used in order to improve the modelling of the spectra. In a change to the previously used cepstral smoothing technique [7], this method is applied in the power spectral domain resulting in a better representation of the energy within the smoothed spectrum.

The smoothed power spectra were frame normalised within a segment and viewed as a probability density function. The EM (Expectation Maximisation) algorithm [1] for finding the maximum likelihood of a mixture model is used to perform the parameter estimation process. In this model the relationship between successive formants is approximated by a linear trajectory through these formants. This provides a parametric representation of the range of possible trajectories for the formant structure of a segment. Any one trajectory is considered to be a Gaussian stochastic process with constant variance whose mean changes as a function of time according to the trajectory.

The following sections describe the main components of this segmental formant vocoder as shown in Figure 1.

2.1. Pitch & Voicing Detection

In order to complete a speech analysis/synthesis system, both pitch estimation and voicing classification are required. In this system, the pitch was estimated using the autocorrelation function. By picking the largest peak above a minimum and below a maximum pitch threshold, an estimate of the pitch is obtained. This was the basis for the smoothing performed in the power spectral domain, in which a raised cosine filter centred at the pitch period is convolved with the power spectrum, suppressing all harmonics above the pitch period.

The voicing decision is made using the autocorrelation function, also in combination with the low-band and high-band energy.

2.2. The Segmental Model

In phonetically segmented speech there are intervals where the temporal variability of the formants is slow. In order to use this temporal correlation, a linear model was formulated. This model requires the estimation of four formants within a phonetic segment of speech with corresponding trajectories. This enables the estimation of any formant frequency at any time frame within that segment. Using segmental Gaussian mixture models with linear trajectories this correlation can be modelled.

Consider a segment S of speech with T frames of frame normalised smoothed power spectrums. Each frame is partitioned into N bins with each bin at time τ within the segment $0 \leq \tau \leq T-1$ having an associated mass $m_x(\tau)$. The probability of a bin number x given a trajectory $t(\tau)$ is defined as

$$p(x|t(\tau)) = p(x|\mu(\tau),\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left[\frac{x-\mu(\tau)}{\sigma}\right]^2} \quad (1)$$

where $\mu(\tau)$ is a linear polynomial of the form

$$\mu(\tau) = \mu + g\tau \quad (2)$$

In this case each mean component μ has an associated linear slope g outlining the trajectory within the segment. Based on this model and given x the log-likelihood of the segment S for the mixture case is formulated as

$$l(S|\Theta) = \sum_{\tau=0}^{T-1}\sum_{x=1}^{N} \log m_x(\tau) \sum_{c=1}^{C} k_c p\left(x|\mu_c(\tau),\sigma_c\right) \quad (3)$$

where Θ denotes the parameter vector $(k_c, \mu_c(\tau), \sigma_c)_{c=1}^{C}$ for C mixtures.

Maximum likelihood parameter estimates $\hat{\Theta}$ may be efficiently computed with the EM algorithm. It involves the iterative application of the following two steps:

- In the E-step, based on current parameter estimates, the posterior probability that the histogram element x at time τ came from track c is estimated as

$$h_c(x,\tau) = \frac{k_c p\left(x|t_c(\tau)\right)}{\sum_{j=1}^{C} k_j p\left(x|t_j(\tau)\right)} \quad (4)$$

- By maximising the auxiliary function with respect to parameters Θ, in the M-step new parameters are estimated using

$$\begin{aligned} \hat{k}_c &= \frac{1}{M}\sum_{\tau=0}^{T-1}\sum_{x=1}^{N}\sum_{c=1}^{C} h_c(x,\tau)m_x(\tau) \\ \hat{\mu}_c &= \frac{\sum_{\tau=0}^{T-1}\sum_{x=1}^{N} h_c(x,\tau)m_x(\tau)\left(x-\hat{g}_c\tau\right)}{\sum_{\tau=0}^{T-1}\sum_{x=1}^{N} h_c(x,\tau)m_x(\tau)} \end{aligned}$$

$$\hat{g}_c = \frac{\sum_{\tau=0}^{T-1} \tau \sum_{x=1}^{N} h_c(x,\tau) m_x(\tau)\,(x - \hat{\mu}_c)}{\sum_{\tau=0}^{T-1} \tau^2 \sum_{x=1}^{N} h_c(x,\tau) m_x(\tau)}$$

$$\hat{\sigma}_c^2 = \frac{\sum_{\tau=0}^{T-1} \sum_{x=1}^{N} h_c(x,\tau) m_x(\tau)\,(x - \hat{\mu}_c(\tau))^2}{\sum_{\tau=0}^{T-1} \sum_{x=1}^{N} h_c(x,\tau) m_x(\tau)}$$

Subsequently $\mu_c(\tau)$ for the duration of the segment is calculated using Eq. 2 and μ_c. The variance of each track is assumed constant across the formants within a segment. Optimal segmentation is achieved using a dynamic programming algorithm based on the log-likelihood. This is describe in the following section.

2.3. Optimal Segmentation

In order to obtain optimal segment boundaries, dynamic programming of the normalised log-likelihood of Eq. 3 is used. A set of log-likelihoods were computed between the current frame and each of the N previous frames. N was limited to a minimum of 3 and a maximum of 12 frames. Note that a transition probability is assigned in order to prevent each segment being the minimum number of frames. This was found to vary slightly between different utterances by different speakers although a constant value was used for all experiments.

Figure 2 shows a spectrogram of an utterance in which four formants are visible. Using this linearly varying mixture of Gaussians technique a set of four means, variances, gradients, and mixture weights along with the segment boundaries were estimated. The boundaries are marked on this diagram and the estimated means superimposed.

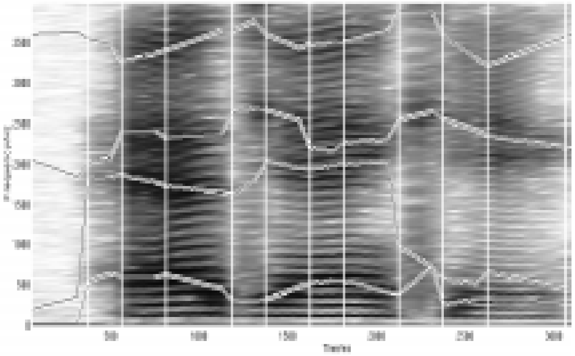

Figure 2: An utterance with segment boundaries and mean trajectories superimposed.

2.4. Vector Quantisation

Standard vector quantisation was utilised in this vocoder. Codebooks were trained, each of dimensionality four (one dimension per Gaussian), except for the mixture weights as they are constrained to sum to one and thus only three of the four parameters are required. 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. In order to avoid domination of bits by the gain and pitch per frame within a segment, a second order polynomial was used to represent the trajectory of gain and a linear polynomial for the pitch, within a segment. The voicing decision within a segment is assumed constant in this implementation of the coder. Table 1 shows these spectral parameter vector conversions for construction of the training data.

The VQ codebooks were built from training data comprising frames of formant analysed natural speech from 50 different speakers lasting approximately 39 minutes in total. Each codebook was trained using the LBG algorithm [4].

Parameter Type	**Representation**	**Bit Allocation**
4 x Mean	Logarithm	10
4 x Std Deviation	Fraction of Mean	6
4 x Gradient	-	8
3 x Weight	Logarithm	7
Voicing	-	1
Power	Logarithm	5
2 x Power Coeffs	-	5
Pitch	Reciprocal	7
Pitch Gradient	-	3
Total bits per Segment		52

Table 1: Conversion of Parameters and Bit Allocation per Segment.

Using a fixed frame rate of 16 ms the operating bit-rate would be 3250 bps. The average frame rate obtained for this segmental vocoder is 7 at which the average operating bit-rate is around 500 bps. This is an 85% reduction in bit-rate.

2.5. The Synthesis System

McAulay described a sinusoidal model for the speech waveform [5], in 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, providing that phase continuity is maintained. The speech waveform can be modelled as a sum of these sine waves. If $s(n)$ represents the sampled speech waveform then

$$s(n) = \sum_i A_i(n) \sin[\phi_i(n)] \qquad (5)$$

where $A_i(n)$ are the amplitudes and $\phi_i(n)$ is the time-varying phase of the i'th partial. 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. Note that the reconstructed phase function is not the same as the original speech waveform but this is perceptually acceptable provided that the magnitude spectrum has been successfully reconstructed.

3. EXPERIMENTAL RESULTS

Figure 3 shows a spectrogram of the utterance *"She had your dark suit in greasy wash water all year"*, and Figure 4 shows the quantised and synthesised version of the same utterance using this segmental vocoder. The operating bit-rate for this sentence is 464 bps and this utterance is available for listening with the proceedings, [sound A0399S01.WAV].

The segmental boundaries play an important role in the speech quality. Various experiments were carried out in order to find the optimum transition probability for obtaining the best segment boundaries. These were judged by eye using the spectrogram with means and frame boundaries superimposed. A problem encountered in specifically unvoiced segments, was that the trajectories of two Gaussians would cross and place a sweeping formant between the boundaries. Also, in some segments the variance was too large, degrading the quality of synthesis.

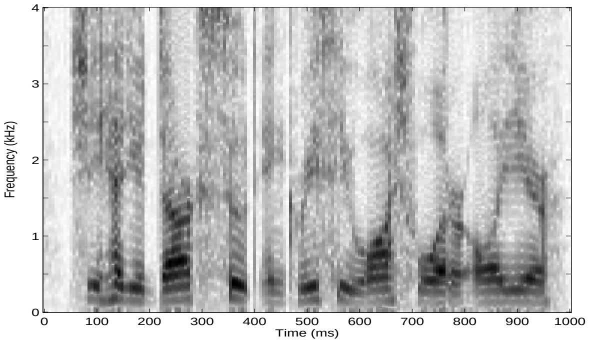

Figure 3: Spectrogram of Original utterance.

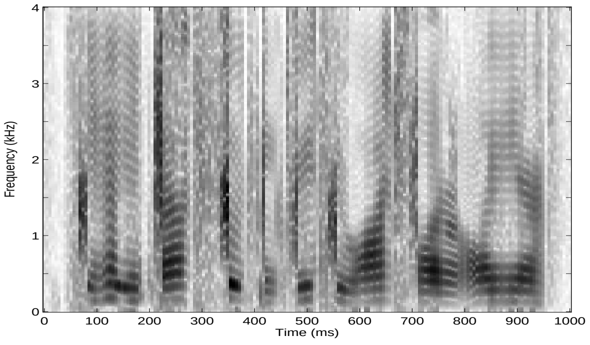

Figure 4: Spectrogram of segmental vocoded utterance.

4. CONCLUSIONS & FURTHER WORK

A segmental format vocoder has been developed using a simple vector quantiser to encode the spectral parameters. An average bit-rate of 500 bps has been obtained. The new smoothing technique showed improved results for formant estimation due to a better representation of spectral power. Further improvements to the segmental model are required in order to place constraints on the gradient of the trajectories to prevent the crossing over of formant tracks. Also further work is to be carried out on improving the search for optimal segment boundaries of the model and tracking the amplitudes and the variance within the segment which would improve the formant structure within the segment. The quality of the synthesised speech is to be improved in order to better model the formant structure represented by the mixtures of Gaussians.

Acknowledgements

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

5. REFERENCES

[1] 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.

[2] W. Goldenthal. *Statistical trajectory models for phonetic recognition.* PhD thesis, Department of Aeronautics and Astronautics, MIT, 1994.

[3] Z. Hu and E. Barnard. Smoothness analysis for trajectory features. *ICASSP*, pages 979–982, April 1997.

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

[5] 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.

[6] S. Roucos, M. Schwartz, and J. Makhoul. A segment vocoder at 150 b/s. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 61–64, Boston, April 1983.

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

[8] P. Zolfaghari and A.J. Robinson. A formant vocoder based on Mixtures of Gaussians. In *Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing*, April 1997.