A MODEL FOR EFFICIENT FORMANT ESTIMATION

L. Welling, H. Ney

Lehrstuhl für Informatik VI, RWTH Aachen – University of Technology
D-52056 Aachen, Germany

ABSTRACT

This paper presents a new method for estimating formant frequencies. The formant model is based on a digital resonator. Each resonator represents a segment of the short–time power spectrum. The complete spectrum is modeled by a set of digital resonators connected in parallel. An algorithm based on dynamic programming produces both the model parameters and segment boundaries that optimally match the spectrum.

The main results of this paper are: 1) Modeling formants by digital resonators allows a reliable estimation of formant frequencies. 2) Digital resonators can be used efficiently in connection with dynamic programming. 3) A recognition test with formant frequencies results in a string error rate of 4.8% on the adult corpus of the TI digit string database.

1. INTRODUCTION

An efficient and compact representation of the time–varying characteristics of speech offers potential benefits for speech recognition. Therefore a variety of approachs such as formant tracking [7, 4, 10], articulatory models [9] and auditory models [5] have been explored. For formant tracking, methods based on linear predictor analysis (LPC) have received considerable attention. Root–finding algorithms are employed to find the zeros of the LPC polynomial or local maxima of the LPC envelope are searched using peak–picking techniques. The problem with root–finding algorithms is that the determination of formant frequencies and bandwidths is only successful for complex–conjugate poles and not for real poles. Peak–picking techniques are vulnerable to merged formants and spurious peaks.

The approach described in this paper avoids the above mentioned problems. In [6], a set of digital formant resonators connected in parallel or in cascade has been proposed for speech synthesis. In this paper, we propose to use digital resonators for formant estimation. We model the power spectrum by K formant models where each model represents one segment of the power spectrum. An algorithm based on dynamic programming produces the set of formant parameters and segment boundaries which optimally match the short–time power spectrum of a speech segment. We have performed recognition tests using formants on the TI digit string database. Formants have also been estimated on the same database in [3].

The paper is organized as follows. Section 2 defines the formant model. Section 3 describes the dynamic programming algorithm that produces the optimal set of segment boundaries. Section 4 contains various experimental results including recognition tests.

2. DEFINITION OF FORMANT MODELS

In this section, we present a model for formant estimation which is based on a set of parallel digital resonators. The frequency range is divided into a fixed number of segments each of which represents a formant. For the moment, the segment boundaries are fixed. In Section 3, we will show how they can be optimized by dynamic programming.

For each segment k with given boundaries, we define a second–order digital resonator. As in general LPC analysis [3, pp. 399], we consider the corresponding predictor polynomial, which is defined as the Fourier transform of the corresponding second–order predictor:

$$A_k(e^{j\omega}) = 1 - \alpha_k e^{j\omega} - \beta_k e^{j2\omega}$$

with predictor coefficients α_k and β_k. $|A_k(e^{j\omega})|^2$ can be written as:

$$\begin{aligned} |A_k(e^{j\omega})|^2 &= 1 + \alpha_k^2 + \beta_k^2 \\ &\quad -2\alpha_k(1-\beta_k)\cos\omega - 2\beta_k\cos 2\omega \\ &= (1+\beta_k)^2 + \alpha_k^2 + \frac{\alpha_k^2(1-\beta_k)^2}{4\beta_k} \\ &\quad -4\beta_k\left[\cos\omega + \frac{\alpha_k(1-\beta_k)}{4\beta_k}\right]^2 \quad . \end{aligned}$$

As can be seen from the above equation, the parameter β_k determines the bandwidth of the resonator. For a resonator, we have the constraint $\beta_k < 0$. $|A_k(e^{j\omega})|^2$ has its global minimum at the resonance or formant frequency φ_k :

$$\varphi_k = \arccos\left(-\frac{\alpha_k(1-\beta_k)}{4\beta_k}\right) \quad . \qquad (1)$$

We denote the beginning point and the end point of segment k by ω_{k-1} and ω_k, respectively. Using the predictor

polynomial, we define the prediction error as follows:

$$E(\omega_{k-1},\omega_k|\alpha_k,\beta_k) = \frac{1}{\pi}\int_{\omega_{k-1}}^{\omega_k} |S(e^{j\omega})|^2\,|A_k(e^{j\omega})|^2 d\omega,$$

where $|S(e^{j\omega})|^2$ denotes the short–time power density spectrum of the speech signal. To find the minimum, we have to optimize the prediction error over α_k and β_k and obtain [3, p. 412]:

$$\alpha_k^{opt} = \frac{r_k(0)r_k(1) - r_k(1)r_k(2)}{r_k(0)^2 - r_k(1)^2}$$

$$\beta_k^{opt} = \frac{r_k(0)r_k(2) - r_k(1)^2}{r_k(0)^2 - r_k(1)^2}$$

with the autocorrelation coefficients $r_k(\nu)$ of segment k for $\nu = 0, 1, 2$:

$$r_k(\nu) := r_{(\omega_{k-1},\omega_k)}(\nu) = \int_{\omega_{k-1}}^{\omega_k} |S(e^{j\omega})|^2 \cos(\nu\omega)d\omega \quad .$$

The minimum error $E_{min}(\omega_{k-1},\omega_k)$ can be expressed as

$$E_{min}(\omega_{k-1},\omega_k) = \min_{\alpha_k,\beta_k} E(\omega_{k-1},\omega_k|\alpha_k,\beta_k) \qquad (2)$$

$$= r_k(0) - \alpha_k^{opt} r_k(1) - \beta_k^{opt} r_k(2) \quad .$$

So far we have considered the prediction error of a single segment k only. We now assume that the whole frequency range is divided into K segments with boundaries $\omega_0 = 0, \ldots, \omega_k, \ldots, \omega_K = \pi$. To define the prediction error for the whole frequency range, we have to sum up the errors of all segments:

$$E = \sum_{k=1}^{K} E_{min}(\omega_{k-1},\omega_k) \quad .$$

A dynamic programming algorithm for finding the optimum segment boundaries is described in Section 3.

In our implementation, the discrete short–time power spectrum $|S(i)|^2$ with discrete frequencies $i = 1, \ldots, 2 \cdot I$ is computed by a $(2 \cdot I)$–point discrete Fourier transform. Segment k ranges from frequency index $(i_{k-1}+1)$ to i_k $(i_0 = 0; i_K = I)$. We calculate the autocorrelation coefficients $r_k(\nu)$ using

$$r_k(\nu) = \sum_{i=i_{k-1}+1}^{i_k} |S(i)|^2 \cos\left(\frac{2\pi\nu i}{2I}\right) \quad .$$

The autocorrelation coefficients can be efficiently computed using the identity

$$r_k(\nu) = T(\nu, i_k) - T(\nu, i_{k-1}) \qquad (3)$$

with look–up tables

$$T(\nu, i) = \sum_{i'=0}^{i} |S(i')|^2 \cos\left(\frac{2\pi\nu i'}{2I}\right) \qquad (4)$$

for $\nu = 0, 1, 2$ and $i = 0, 1, \ldots, I$.

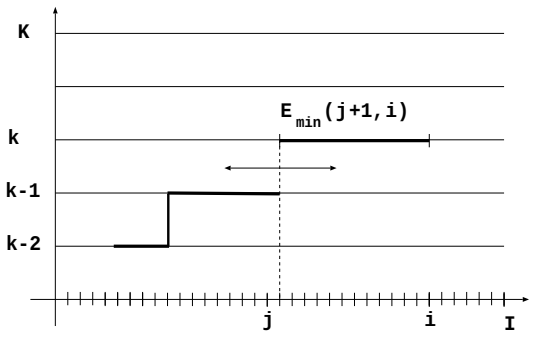

Figure 1: Segmentation by dynamic programming.

3. DYNAMIC PROGRAMMING ALGORITHM

The task is now to find the segment boundaries $i_1, \ldots, i_{K-1}$ so that

$$\sum_{k=1}^{K} E_{min}(i_{k-1}+1, i_k)$$

is minimized $(i_0 = 0, i_K = I)$. Dynamic programming [1, 2] provides an efficient solution.

We introduce an auxiliary quantity $F(k,i)$ which is defined as the error of the best segmentation of the frequency interval $[1,i]$ into k segments. By decomposing the frequency interval $[1,i]$ into two frequency intervals $[1,j]$ and $[j+1,i]$ and using the optimality in the definition of $F(k,i)$, we obtain the recurrence relation of dynamic programming:

$$F(k,i) = \min_j [F(k-1,j) + E_{min}(j+1,i)] \quad . \qquad (5)$$

As Equation (5) shows, the best segmentation of the frequency interval $[1,j]$ into $k-1$ segments is utilized to determine the partition of the frequency interval $[1,i]$ into k segments. Figure 1 gives an illustration of the dynamic programming Equation (5).

The optimum segment boundaries are obtained along with the minimum error $F(K,I)$ by recursively applying Equation (5). Table 1 summarizes the complete algorithm. The first step is to fill the look–up tables defined by Equation (4). Then the values of $E_{min}(j+1,i)$ for $0 < j+1 < i$ and $1 \le i \le I$ are calculated using Equations (2) and (3). The algorithm employs a backpointer array B to obtain the segment boundaries. After the segmentation process, the formant frequencies for each segment are calculated by Equation (1). The number of operations in the inner loop of the algorithm is $K \cdot I^2/2$.

4. EXPERIMENTAL RESULTS

4.1. Formant Estimation

We have tested the formant model on the TI digit string database. First we perform a signal pre–emphasis by calculating the first–order difference of the sampled speech

Table 1: Dynamic programming algorithm for finding the segment boundaries.

initialisation: compute $E_{min}(j,i)$ for $j < i$
for each frequency i from 1 to I do
for each segment k from 1 to K do
$F(k,i) = \infty$
for each frequency j from 1 to $i-1$ do
if $F(k-1,j) + E_{min}(j+1,i) < F(k,i)$
$F(k,i) = F(k-1,j) + E_{min}(j+1,i)$
$B(k,i) = j$
traceback: $i(K) = I$
for each segment k from K to 1 do
$i(k-1) = B(k,i(k))$
calculate $\varphi(k)$ and $\beta(k)$

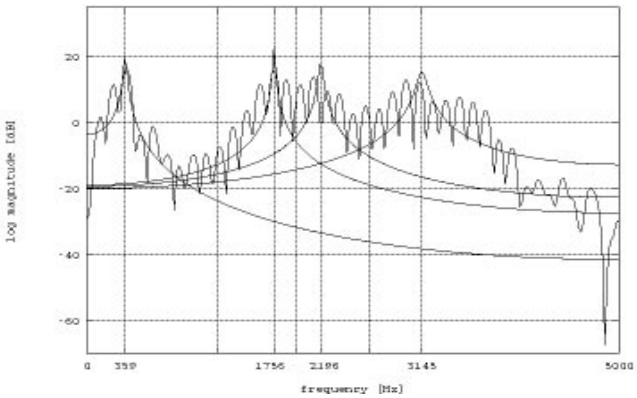

Figure 2: Spectrum and formant models for frame 95 of digit string *73* by male talker *AH*.

signal. Every 10 ms, a 20–ms Hamming window is applied to overlapping speech segments and the short–time power spectrum is computed by a 1024–point fast Fourier transform. The frequency range from 0 to 5 kHz is used. We fix the number of formants to $K = 4$. There is no smoothing of the formant frequencies.

Figure 2 depicts the short–time power spectrum of a speech frame from the 'ee' sound in 'three' together with the formant models that were obtained. In Figure 2, the four formant frequencies and the segment boundaries are represented by vertical lines. The bandwidths of the formants are relatively small, which we attribute to the segmentwise formant definition.

Figure 3 shows the spectrogram (after local energy normalisation) of the digit 'three' together with the estimated formant frequencies. There is a good agreement between the formant frequencies and the spectrogram.

Figure 4 shows a histogram for each of the four formant

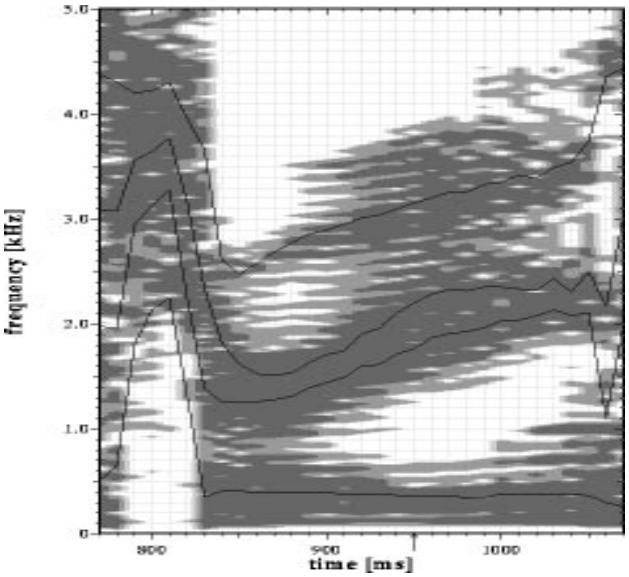

Figure 3: Spectrogram and formant contours of the word 'three' (frames 77-107 of digit string *73* by talker *AH*; the arrow marks frame 95 shown in Figure 2).

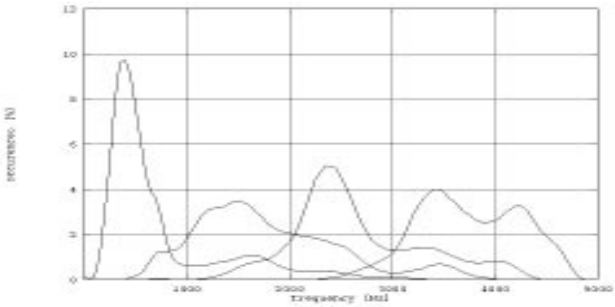

Figure 4: Histogram of formant frequencies over male training speakers of TI digit string database.

frequencies that were generated using the utterances of the male training speakers of the TI digit string database. For the histogram, the silence frames in the acoustic signal were omitted.

4.2. Formant–Based Recognition

The estimated formant frequencies were used to form the acoustic vectors for recognition experiments on the TI digit string database. The recognition system is based on hidden Markov models with continuous observation densities. Its characteristic features are [11]: 1) gender–dependent word models for 11 English digits including '*oh*' and gender–dependent silence models; 2) 357 states plus 1 state for silence per gender; 3) single Laplacian densities with state dependent deviation vectors; 4) maximum likelihood training in the Viterbi approximation.

The signal analysis is performed every 10 ms. For each time frame t, the acoustic vector $y(t)$ consists of signal

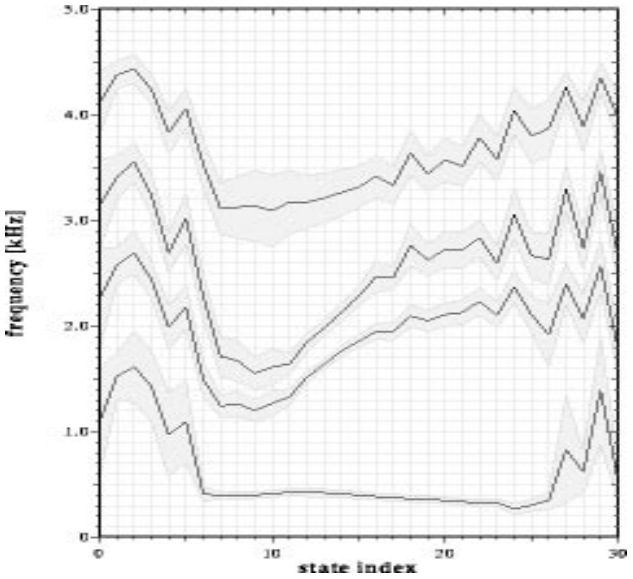

Figure 5: Formant–based reference for 'three'.

energy and four formant frequencies. This vector $y(t)$ is augmented by first–order derivatives. The resulting acoustic vector is $x(t) = [y(t), y(t) - y(t - \Delta t)]^T$ with $\Delta t = 3 \cdot 10$ ms.

Figure 5 shows the male reference model for the digit 'three' that was obtained by training. In addition to the formant frequency contours, Figure 5 also shows the absolute deviation of the Laplacian models represented by a gray stripe.

Table 2 gives recognition results for both formants and cepstrum on the TI digit string database. For both types of acoustic vectors, Table 2 shows the word error rates, string error rates and the number of components of the acoustic vector. For the cepstrum, the acoustic vector consists of 16 cepstral coefficients, 16 first–order derivatives and 16 second–order derivatives [11]. As can be seen in Table 2, promising recognition results were obtained for the formants. In particular, it should be noted that the number of components of the acoustic vector is significantly smaller for the formants than for the cepstrum. Furthermore, there was no smoothing or other postprocessing of the formant trajectories. To the best of our knowledge, this is one of the few recognition systems that are based solely on formant contours. Considering that this work started only recently, we see room for further improvements in formant-based speech recognition in the future.

Table 2: Recognition errors for formants and cepstrum (TI digit string database).

Acoustic vector	Number of components	Word error rate [%]	String error rate [%]
Formants	2·5	1.7	4.8
Cepstrum	3·16	0.6	1.8

5. CONCLUSIONS

This paper has presented a new approach to formant estimation: 1) The short–time power spectrum is decomposed into segments each of which is modeled by a digital resonator. 2) The segment boundaries are optimised by dynamic programming.

The estimated formant frequencies have been analysed using spectrograms and histograms. In a recognition test on the adult corpus of the TI digit string database, a string error rate of 4.8% has been achieved with four formant frequencies and signal energy.

REFERENCES

1. R. Bellman, S. Dreyfus, "Applied Dynamic Programming," Princeton Univ., Princeton, NJ, 1962.
2. J. S. Bridle, N. C. Sedgewick, "A method for segmenting acoustic patterns, with applications to automatic speech recognition," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Hartford, CN, pp. 656-659, May 1977.
3. M. A. Bush, G. E. Kopec, "Network-based connected digit recognition," *IEEE Trans. on Acoust., Speech and Signal Processing*, Vol. ASSP-35, No. 10, pp. 1401-1413, October 1987.
4. A. Crowe, M. A. Jack, "Globally optimising formant tracker using generalised centroids," *Electronics Letters*, Vol. 23, No. 19, pp. 1019-1020, September 1987.
5. M. J. Hunt, C. Lefèbvre, "Speaker dependent and independent speech recognition experiments with an auditory model," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, New York, pp. 215-218, April 1988.
6. D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, Vol. 67, No. 3, pp. 970-995, March 1980.
7. G. E. Kopec, "Formant tracking using hidden Markov models and vector quantization," *IEEE Trans. on Acoust., Speech, Signal Processing*, Vol. ASSP-34, No. 4, pp. 709-729, August 1986.
8. L. Rabiner, R.-W. Schafer, "Digital processing of speech signals," Prentice-Hall, Englewood Cliffs, NJ, 1978.
9. H. B. Richards, J. S. Mason, M. J. Hunt, J. S. Bridle, "Deriving articulatory representations of speech," *Proc. Europ. Conf. on Speech Communication and Technology*, Madrid, pp. 761-764, September 1995.
10. R. C. Snell, F. Milinazzo, "Formant location from LPC analysis data," *IEEE Trans. on Speech and Audio Processing*, Vol. 1, No. 2, pp. 129-134, April 1993.
11. L. Welling, H. Ney, A. Eiden, C. Forbrig, "Connected digit recognition using statistical template matching," *Proc. Europ. Conf. on Speech Communication and Technology*, Madrid, pp. 1483-1486, September 1995.