FORMANT FREQUENCY ESTIMATION IN NOISE

*Bin Chen and Philipos C. Loizou**

University of Texas at Dallas, Dept. of Electrical Engineering
Richardson, TX 75083
*loizou@utdallas.edu

ABSTRACT

This paper addresses the problem of formant frequency estimation of speech signals corrupted by colored noise. The spectrum is sequentially segmented inot K segments so that each segment contains a single formant. A segmentation metric based on Wiener filter theory is proposed for determining the segment boundaries. A peak-picking algorithm is used for estimating the formant frequencies in each segment. Results obtained using vowels embedded in +5 dB S/N speech-shaped noise, indicated that the proposed algorithm produced formant frequencies which were comparable to those estimated in quiet.

1. INTRODUCTION

Apart from a variety of formant tracking approaches [1][2], considerable attention has been paid to methods based on linear prediction analysis (LPC) [3][4]. However, capturing and tracking formants accurately from noisy speech is not easy, largely because the accuracy of root-finding algorithms based on LPC is sensitive to the noise level.

In [5][6], a set of parallel digital formant resonators has been proposed for speech synthesis or formant frequency estimation. In this paper, we propose the use of a sequential digital resonator model for spectrum segmentation. The spectrum segmentation is implemented sequentially from low to high frequencies. For each spectral segment, a digital resonator is first determined to represent the spectral segment. A metric based on Wiener filter theory is proposed to determine the segment boundaries. After identifying the spectral segments containing the formants, we apply a peak-picking algorithm on each spectral segment to find the formant frequency. This approach was taken since the LPC-based digital resonators are sensitive to the noise level. A major advantage of the proposed method is that it determines the segment boundaries sequentially and avoids the need for dynamic programming as done in [5] and [7].

This paper is organized as follows. Section 2 describes the formant estimation model, Section 3 presents the proposed formant frequency estimation algorithm, Section 4 presents the experimental results, and Section 5 gives the conclusions.

2. FORMANT ESTIMATION MODEL

In this section, a model is described for formant estimation that is implemented using a set of digital resonators. Each resonator represents a formant in a segment in the frequency domain. The spectrum is divided into segments such that only one formant resides in each segment. For the convenience of representing the digital resonator, the segment boundaries are assumed to be fixed. In the next section, we show how to determine the segment boundaries sequentially using a Wiener-based metric.

Each formant in a spectral segment k is represented by a second-order prediction filter. The second-order prediction filter for the formant in the spectral segment k is given by the all-pole model $1/A_k(z) = 1/(1+a_k z^{-1}+\beta_k z^{-2})$. The formants can be considered as being generated by a second-order system driven by white noise. $A_k(z)$ is a whitening filter that whitens the formant spectrum, i.e., it flattens the spectrum in segment k. If $A_k(z)$ is used as a notch filter, it will notch the corresponding formant out of the spectrum. In our application, we adopt the notch filter definition in [8]

$$H_k(z) = \gamma(1 + \alpha z^{-1} + \beta z^{-2}) \qquad (1)$$

where $\alpha = e^{-2\pi B}$, $\beta = -2e^{-\pi B}\cos\omega$ and $\gamma = 1/(1+\alpha+\beta)$ are specified by the notch frequency ω and the bandwidth B. Note that $A_k(z)$ is similar to $H_k(z)$ except for the scalar γ. Thus, we can find the notch filter by determining the segmental system transfer function $H_k(z)$.

According to [5], the optimum prediction coefficients of the notch filter are given by:

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

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

where $r_k(m)$ are the autocorrelation coefficients obtained for segment k

$$r_k(m) = r_{(w_{k-1},w_k)}(m) = \frac{1}{\pi}\int_{\omega_{k-1}}^{\omega_k} |S(e^{j\omega})|^2 \cos(m\omega) d\omega \tag{3}$$

Substituting α_k^{opt} and β_k^{opt} obtained above to Eq. (1) gives us the desired notch filter $H_k(z)$ of the kth band in the spectrum. The scalar γ is independent of minimization of the prediction error and is determined after the α_k^{opt} and β_k^{opt} are found.

As in [5], we use a discrete approximation of the integral in Eq.(3). The frequency range $[0\ \pi]$ is divided into I equally spaced intervals $\Delta\omega\ (= \pi/I)$ with grid $\pi i/I$, $i = 0, 1, ..., I$. Therefore, the segment boundaries $\omega_0 = 0, ..., \omega_k, ..., \omega_K = \pi$ are replaced by the indices $i_0 = 0, ..., i_k, ..., i_K = I$, and $r_k(m)$ is given by

$$r_k(m) = \frac{1}{I}\sum_{i=i_{k-1}}^{i_k} |S(i)|^2 \cos\left(\frac{2\pi m i}{2I}\right) \tag{4}$$

with $S(i) = S(\omega)|_{\omega=e^{j(2\pi i/2I)}}$.The above autocorrelation sequence is determined for a specific spectral segment $[i_{k-1}\ i_k]$, and is expected to vary accordingly with the spectral segment. Experiments showed that the autocorrelation sequence does not change much when a strong formant dominates the spectral segment even after the spectral segment is expanded to include a second formant.

3. PROPOSED FORMANT FREQUENCY ESTIMATION ALGORITHM IN NOISE

So far we described a formant frequency model for a single spectral segment k. That is, we assumed that the segment boundaries were known. In this section, we propose a segmentation metric, motivated by Wiener filter theory, that identifies the boundaries of the K segments of the spectrum containing the K formants.

Suppose that the input to a Wiener filter is a signal with an additive noise, i.e., $x(n) = s(n) + n(n)$, and the desired signal is the noise, i.e., $d(n) = n(n)$. From the orthogonality principle, we know that

$$\begin{aligned} & E[e(n)x(n-l)] = 0 \\ = \quad & r_{nn}(l) - \sum_{k=0}^{\infty} h_w(k) r_{xx}(l-k) \end{aligned} \tag{5}$$

where $h_w(n)$ is the Wiener filter, $e(n)$ is the estimation error, and $r_{nn}(l)$ and $r_{xx}(l)$ are the autocorrelation sequences of the noise and noisy speech signal respectively. For a given notch filter $h(n)$, we can produce the prediction residual $w(n)$ of the clean signal as

$$w(n) = \sum_{k=0}^{M-1} h(k) s(n-k) \tag{6}$$

where $h(0) = 1$, and $M = 3$. Now, if we replace the Wiener filter $h_w(n)$ in Eq.(5) with the notch filter $h(n)$, we get:

$$\begin{aligned} & E[e(n)x(n-l)] \\ = \quad & r_{nn}(l) - \sum_{k=0}^{M-1} h(k) r_{xx}(l-k) \end{aligned} \tag{7a}$$

Since $x(n) = s(n) + n(n)$, we get from Eq. (7a):

$$\begin{aligned} & E[e(n)x(n-l)] = \\ & r_{nn}(l) - E[w(n)x(n-l)] - \sum_{k=0}^{M-1} h(k) r_{nn}(l-k) \neq 0 \end{aligned} \tag{7b}$$

Note that Eq. (7b) is no longer equal to zero since the notch filter $h(n)$ in Eq.(7b) is not the optimum Wiener filter. Since the prediction residual $w(n)$ is independent of the noisy signal $x(n)$, the second term $E[w(n)x(n-l)]$ in Eq.(7b) ought to be zero. In practice, however, $w(n)$ becomes white only if $h(n)$ whitens $s(n)$. As the upper boundary of a segment expands, the notch filter $h(n)$ will gradually become more and more matched with the formant in the segment, and $E[w(n)x(n-l)]$ will become smaller and smaller. When $E[w(n)x(n-l)]$ reaches its minimum, or $E[e(n)x(n-l)]$ attains its maximum, the whole formant will be matched and contained in the segment. As mentioned earlier, the notch filter $h(n)$ will not change much even if the next formant is included. That is, $E[e(n)x(n-l)]$ reaches a maximum and saturates thereafter. The point at which the maximum is reached is indicative of a segment boundary. We therefore use the energy of $E[e(n)x(n-l)]$ as the segmentation metric.

The third term $\sum h(k) r_{nn}(l-k)$ in Eq.(7b) may also become small as $h(n)$ changes. In order to offset the effect of this undesired term, we add the term $\sum h(k) r_{nn}(l-k)$ in Eq.(7a). The final segmentation metric then becomes:

$$\begin{aligned} & E_k[e(n)x(n-l)] \\ = \quad & r_{nn}^k(l) - \sum_{m=0}^{M-1} h_k(m) r_{xx}^k(l-m) \\ & + \sum_{m=0}^{M-1} h_k(m) r_{nn}^k(l-m) \end{aligned} \tag{8}$$

where $h_k(m)$ and $r^k(m)$ represent the notch filter and the autocorrelation sequence calculated from the kth spectral segment $[\omega_{k-1}\ \omega_k]$ respectively. The energy of $E_k[e(n)x(n-l)]$ is used as the segmentation metric and is denoted by

$$E_{ex}(\omega_{k-1}, \omega_k) = \sum_{l=0}^{M-1} E_k[e(n)x(n-l)] \tag{9}$$

The metric saturation point, which is also the segment boundary point, is defined to be the point at which the following condition is satisfied:

$$\left| \frac{E_{ex}(w_{k+m}) - E_{ex}(w_k)}{E_{ex}(w_k)} \right| < \varepsilon \qquad (10)$$

where $E_{ex}(w_k)$ denotes $E_{ex}(\omega_{k-1}, \omega_k)$ for simplicity. The delay index m is used to ensure that there is a long enough saturation period before a true saturation point is detected. Empirically, m should be selected such that the saturation period is no less than 300 Hz. The constant ε is empirically determined. Figure 1 shows an example of the segmentation of a noisy vowel spectrum.

Once the segmentation of the formant region is determined, we considered peak-picking the spectrum. The basic idea is to segment the noise spectrum to have only one formant in each segment, and then for each segment, peak-pick the spectrum to get an estimate for the formant frequency of the noisy speech spectrum.

The above segmentation algorithm requires access to the autocorrelation sequence of the clean signal, which we do not have. To estimate the clean autocorrelation sequence, we considered pre-processing the signal by the spectral subtraction algorithm [9] to get an estimate of the enhanced signal spectrum. The autocorrelation sequence is obtained using Eq.(4) but with $S(i)$ being replaced with the enhanced speech spectrum.

3.1. Proposed Algorithm

The proposed algorithm is outlined below:

```
Initialization:
    $k = 1$; $i_{k-1} = 0$; $i_k = 1$;
    $K =$ desired number of formants
Step 1. Loop (for segment $k$):
    (1) Calculate $r_{xx}^{(i_k)}(l)$ and $r_{nn}^{(i_k)}(l)$ using Eq. (4)
    (2) Use Equations (2a), (2b) and (4) to calculate the notch filter $h^{(i_k)}(n)$
    (3) Use Equations (8) and (9) to estimate $E_{ex}(\omega_{k-1}, \omega_k)$
    (4) if $E_{ex}(\omega_{k-1}, \omega_k)$ reaches a saturation point (according to Eq. 10), then:
            $k$th boundary = $i_k$
            Peak-pick spectrum to estimate formant frequency.
            go to Step 2
        end
    (5) $i_k = i_k + 1$
End
Step 2. $k = k + 1$
    $i_{k-1} = i_k$
    if $k > K$, stop
    else, go to Step 1
```

In our implementation, the autocorrelation sequence of the noise, $r_{nn}(l)$, was estimated using the first few speech-absent frames of the noisy speech signal. The speech signal was processed using 10-ms duration Hamming windows with 50% overlap between adjacent frames, and the spectrum in Eq. 4 was obtained using the FFT.

	/oo/		/ah/		/ey/		/iy/	
	LPC	SEF	LPC	SEF	LPC	SEF	LPC	SEF
F_1	14.8	50.7	11.8	41.2	11.0	47.9	13.9	44.2
F_2	18.8	81.7	15.6	63.7	16.2	45.7	9.5	89.6
F_3	25.5	137	21.8	156	21.9	83.4	18.5	69.9

Table 1. Standard deviations (Hz) of formant frequency errors for synthetic vowels using the proposed algorithm (SEF) and the LPC algorithm. The formant frequencies of the LPC algorithm were obtained in quiet, while the frequencies of the SEF algorithm were based on vowels embedded in +5 dB speech-shaped noise.

4. EXPERIMENTAL RESULTS

The proposed formant frequency estimation algorithm was evaluated using real and synthetic vowels. Four natural vowels, /u/, /a/, /ei/ and /i/, corrupted by speech-shaped noise at +5 dB S/N were used for evaluation. The vowels were contained in the words “hood”, “hod”, “hayed” and “heed” and were produced by a male speaker. The estimated formant tracks are shown in Figure 2. For comparative purposes, we also estimated the formant frequencies of these vowels *in quiet* using two other methods based on LPC(16th order) and dynamic programming [5]. As can be seen, our estimated formant frequencies are comparable to the estimated formant frequencies in quiet.

The same vowels were also synthesized using the Klatt synthesizer [6], and corrupted by a +5dB speech-shaped noise. Each test consisted of 200 trials in which the F1 was varied ±200 Hz and the F2 and F3 frequencies were varied ±150 Hz around the center of the corresponding formant frequencies. Standard derivations were measured of the differences between the true formant frequencies and the estimated formant frequencies. The results are tabulated in Table 1. For comparative purposes, we also list the standard deviations of the formant frequencies of the same vowels estimated in quiet using the LPC method. Results indicated that the estimation of the F1 frequency was more accurate than the estimation of the F2 and F3 frequencies.

5. SUMMARY AND CONCLUSIONS

A new method for estimating formant frequencies in noise was proposed based on sequential determination of spectral segments and formant frequencies. The spectrum was

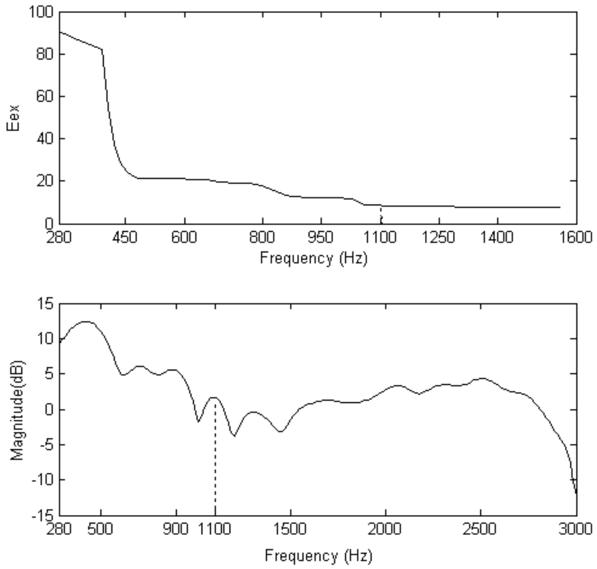

Fig. 1. The top panel shows values of the segmentation metric as a function of frequency. Saturation point was estimated to be 1100 Hz. Bottom panel shows the noisy spectrum of the vowel /ey/. In this example, the F1 region was determined to be 0-1100 Hz.

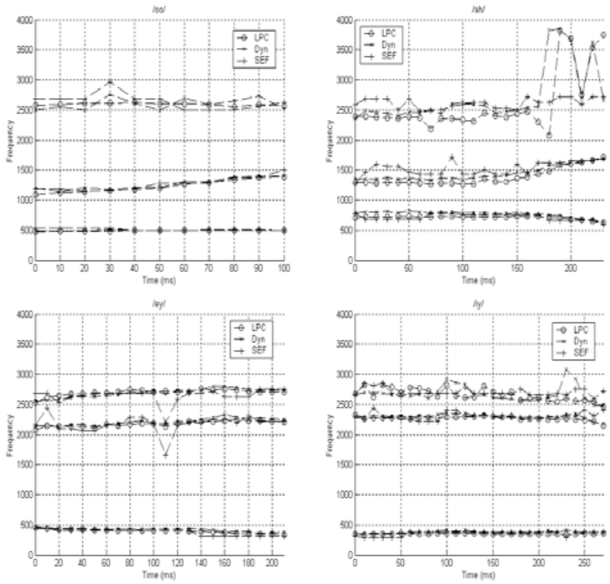

Fig. 2. Formant tracks for four vowels in +5 dB S/N estimated using the proposed formant frequency estimation algorithm (SEF). For comparison, we superimpose the formant tracks of the vowels estimated in quiet by the LPC and dynamic programming based algorithms (Dyn) [5].

sequentially segmented into K segments using a new segmentation metric based on Wiener filter theory. No specific assumptions were required for the statistics of the noise. Experimental results showed that the estimated formant frequencies of vowels embedded in +5 dB speech-shaped noise were comparable to the formant frequencies estimated in quiet.

6. REFERENCES

[1] A. Crowe and M.A.Jack, ”Globally optimizing formant tracker using generalized centroids,” *Electron. Lett.*, vol. 23, pp. 1019-1020, Sept. 1987.

[2] G. E. Kopec, ”Formant tracking using hidden Markov models and vector quantization,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 709-729, Aug. 1986.

[3] S. McCandless, ”An algorithm for automatic formant extraction using linear prediction spectra,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 135-141, 1974.

[4] R. C. Snell and F. Milinazzo, ”Formant location from LPC analysis data,” *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 129-134, Apr. 1993.

[5] L. Welling and Hermann Ney, ”Formant Estimation for Speech Recognition,” *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 36-48, Jan. 1998.

[6] D. H. Klatt, ”Software for a cascade/parallel formant synthesizer,” *J. Acoust. Soc. Amer.*, vol. 67, pp. 970-995, Mar. 1980.

[7] H. S. Chhatwal and A. G. Constantinides, ”Speech spectral segmentation for spectral estimation and formant modeling,” in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Dallas, TX, apr. 1987, pp. 316-319

[8] A. Watanabe, ”Formant Estimation Method Using Inverse-Filter Control,” *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 317-326, May 2001.

[9] M. Berouti, R. Schwartz and J. Makhoul, ”Enhancement of speech corrupted by acoustic noise”, *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 4, pp. 208-211, Apr 1979