Robust Formant Tracking for Continuous Speech with Speaker Variability

Kamran Mustafa and Ian C. Bruce, *Member, IEEE*

***Abstract*— Several algorithms have been developed for tracking formant frequency trajectories of speech signals, however most of these algorithms are either not robust in real-life noise environments or are not suitable for real-time implementation. The algorithm presented in this paper obtains formant frequency estimates from voiced segments of continuous speech by using a time-varying adaptive filterbank to track individual formant frequencies. The formant tracker incorporates an adaptive voicing detector and a gender detector for formant extraction from continuous speech, for both male and female speakers. The algorithm has a low signal delay and provides smooth and accurate estimates for the first four formant frequencies at moderate and high signal-to-noise ratios. Thorough testing of the algorithm has shown that it is robust over a wide range of signal-to-noise ratios for various types of background noises.**

***Index Terms*— Formant estimation, speech analysis, hearing aids, speech enhancement.**

I. Introduction

FORMANT frequency trajectories are major acoustical cues for the identification of phonemes including vowels [1]–[4], nasal consonants [5], diphthongs [6], and consonants in consonant-vowel transitions [7]. Sound-induced hearing loss can cause cochlear hair cell damage, leading to the degradation of the auditory nerve response to formant frequencies [8], [9]. It is likely that these degradations contribute to decreased intelligibility of speech for people suffering from sound-induced hearing loss. Hearing aids that apply amplification independently across different frequency bands probably cannot compensate satisfactorily for this type of hearing loss [9], [10]. However, an amplification scheme for hearing aids called Contrast Enhanced Frequency Shaping (CEFS) that should improve speech perception has been proposed by Miller and colleagues [11]. CEFS takes into account across-frequency distortions introduced by the impaired ear and requires robust formant frequency estimates to allow dynamic, speech-spectrum-dependent amplification of speech in hearing aids [9], [11], [12]. The accurate and robust formant frequency estimation required for CEFS is not easy to accomplish in real-time for continuous speech. This task becomes even more difficult in real-life noise environments and with speaker variability (i.e., in cases where the speaker is unknown, the gender of the speaker is unknown, or different individuals are speaking at different times or simultaneously).

Traditional formant frequency estimation methods are based on spectral analysis and 'peak picking' techniques (e.g., [13]–[17]). However, a comparative analysis of some of these algorithms has shown that they are neither accurate nor robust in transient background noise [18]. A more reliable formant estimation technique has been proposed by Rao and Kumaresan [19]. This method is based on pre-filtering speech using a time-varying adaptive filter for each formant before spectral peak estimation. The pre-filtering limits the spectral region of estimation for each formant frequency and thereby minimizes the effects of the neighboring formants or background noise on the estimates. The Rao and Kumaresan approach provides reasonably accurate formant frequency estimates for strongly-voiced segments of speech. However, the algorithm is not robust, does not recover well after a period of silence, and is unreliable during unvoiced speech segments. These factors make the Rao and Kumaresan algorithm unsuitable for implementation in hearing aids for CEFS amplification of continuous speech [18]. Bruce and colleagues [20] proposed various improvements to the Rao and Kumaresan algorithm to overcome these limitations. The Bruce *et al.* algorithm includes a formant energy detector and a voicing detector, so that the algorithm does not track formants during unvoiced speech segments, during periods of silence, or when a formant has insufficient energy for reliable spectral estimation [20].

The formant tracking algorithm presented in this paper has several improvements upon the Bruce *et al.* scheme to make it more robust in continuous speech, to speaker variability, and to different types of background noises present in real-life environments. In Section II we describe the improvements that have been made to the Bruce *et al.* algorithm, in Section III we describe the results of quantitative and qualitative testing of the new algorithm in a variety of background noise conditions, and in Section IV we discuss the implications of these results and draw some conclusions about the utility of the algorithm for a range of speech processing applications. Preliminary results have been presented in [18] and [21]. MATLAB source code for the algorithm is available on request.

II. The Formant Tracking Algorithm

A block diagram showing the main features of the formant tracker is shown in Fig. 1. First, the natural spectral tilt of the signal [14], [22] is removed via a highpass pre-emphasis filter (2^{nd}-order Butterworth filter with a cutoff frequency of 3 kHz). An analytic version of the pre-emphasized speech signal is then calculated using an approximate Hilbert transformer

This work was supported by the National Science and Engineering Research Council, Canada, (Discovery Grant 261736) and the Barber-Gennum Endowed Chair in Information Technology.

K. Mustafa was with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, Ontario L8S 4K1, Canada. He is now with Nortel Networks, Ottawa, Ontario K2H 8E9, Canada.

I. Bruce is with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, Ontario L8S 4K1, Canada (e-mail: ibruce@ieee.org).

(20th-order linear-phase FIR filter [23]). The primary reason behind converting the real-valued signal into its analytic representation is to allow the use of complex-valued filters in the formant filterbank (see Section II-A below). The conversion also decreases the amount of aliasing in the signal, increasing the accuracy of the spectral estimation technique used for formant frequency estimation [24].

The algorithm was implemented for speech signals sampled at $F_s = 8$ kHz and with an RMS energy of 0 dB over the entire duration of the signal.

A. Adaptive Bandpass Filterbank

The adaptive bandpass filterbank used in the formant tracking algorithm is similar to the one proposed by Rao and Kumaresan [19] but has been modified to further suppress the effects of the pitch from the first-formant estimation. Each band of the filterbank consists of an all-zero filter (AZF) cascaded with a single-pole dynamic tracking filter (DTF). The combination of the AZF and the DTF is called a *formant filter* and is responsible for bandpass filtering the speech signal prior to estimating individual formant frequencies. Complex filters are used to simplify normalization of the filter frequency response to give unity gain and zero phase lag at the filter center frequency [19]. The zeros and pole of each formant filter are updated over time, based on the previous formant frequency estimates, allowing dynamic suppression of interference from neighboring formants and from background noise sources, while tracking an individual formant frequency as it varies with time.

In the filters for tracking $F2$–$F4$, each AZF has three zeros that are set to the previous value of the other three formant frequency estimates (obtained from the other three formant trackers). The transfer function of the k^{th} AZF (where $k = 2, 3$ or 4) at time n is:

$$H_{\text{AZF}k}[n,z] = K_k[n] \times \prod_{\substack{l=1 \\ l \neq k}}^{4} (1 - r_z e^{j2\pi Fl[n-1]} z^{-1}), \quad (1)$$

where

$$K_k[n] = \frac{1}{\prod_{\substack{l=1 \\ l \neq k}}^{4} (1 - r_z e^{j2\pi (Fl[n-1] - Fk[n-1])})} \quad (2)$$

and r_z is the radius of the zeros on the z-plane, $Fl\,[n-1]$ and $Fk\,[n-1]$ are the formant frequency estimates at the previous time index from the l^{th} and k^{th} formant trackers, respectively. The term $K_k[n]$ ensures that the AZF has unity gain and zero phase lag at the estimated formant frequency of the k^{th} component. A value of $r_z = 0.98$ is used [19].

The transfer function of the 1st-formant AZF is slightly different than those of the other three AZFs. The AZF of the $F1$ filter has an additional zero at the pitch frequency $F0$ to prevent the $F1$ tracker from erroneously tracking the signal energy at the $F0$ frequency. Therefore, the transfer function of the 1st AZF at index n is:

$$H_{\text{AZF1}}[n,z] = K_1[n] \times \prod_{\substack{l=0 \\ l \neq 1}}^{4} (1 - r_z e^{j2\pi Fl[n-1]} z^{-1}), \quad (3)$$

where

$$K_1[n] = \frac{1}{\prod_{\substack{l=0 \\ l \neq 1}}^{4} (1 - r_z e^{j2\pi (Fl[n-1] - F1[n-1])})} \quad (4)$$

and $F0[n-1]$ is the pitch estimate at the previous time index, which is provided to the 1st-formant filter by the gender detector described in Section II-D below. Note that $F0$ is only estimated during voiced speech, so during unvoiced segments of speech the moving average value of the $F0$ estimates is used.

The DTF in each formant filter is made up of a single pole, where the location of the pole is always set to the previous estimate of the formant frequency of that formant filter. The transfer function of the k^{th} DTF at time n is:

$$H_{\text{DTF}k}[n,z] = \frac{1 - r_p}{(1 - r_p e^{j2\pi Fk[n-1]} z^{-1})}, \quad (5)$$

where r_p is the radius of the pole and $Fk[n-1]$ is the formant estimate of the k^{th} formant tracker at the previous time index. A value of $r_p = 0.90$ is used [19].

After the placement of the pole and zeros for each formant filter, the complex filter coefficients of the four formant filters are calculated. These filter coefficients are then used to filter the analytic speech signal into four band-limited spectral regions from which the four formant frequencies are estimated. Example frequency responses of the four formant filters at a particular time are shown in Fig. 2. The positions of the pole and the zeros of each formant filter are updated every sample, as the formant frequency estimates vary with time. All four formant filters have unity gain and zero phase lag at the location of the pole, i.e., at the peak of the bandpass filter that corresponds to the previous estimate of the formant frequency.

Figure 3 shows spectrograms of the original speech signal and the signal from each of the four formant filters after the original signal has been adaptively filtered, using the formant filterbank. It can be seen that in each of the filtered signals the energy of the neighboring formants is greatly reduced and each filter output contains energy primarily from only one formant. The energy at the pitch frequency is also attenuated in the output of the $F1$ formant filter.

B. Spectral Estimation via Linear Prediction

The first four formant frequencies of voiced speech segments are estimated from the four bands of the adaptive bandpass filterbank using first-order linear prediction on each band. The analytic signal from each band is first windowed using a 20-ms Hamming window. Next, a single linear predictive coefficient (LPC) of the windowed frame is calculated for each band using the autocorrelation method (e.g., [25], [26]), fitting a single-pole model to the windowed signal in each band.

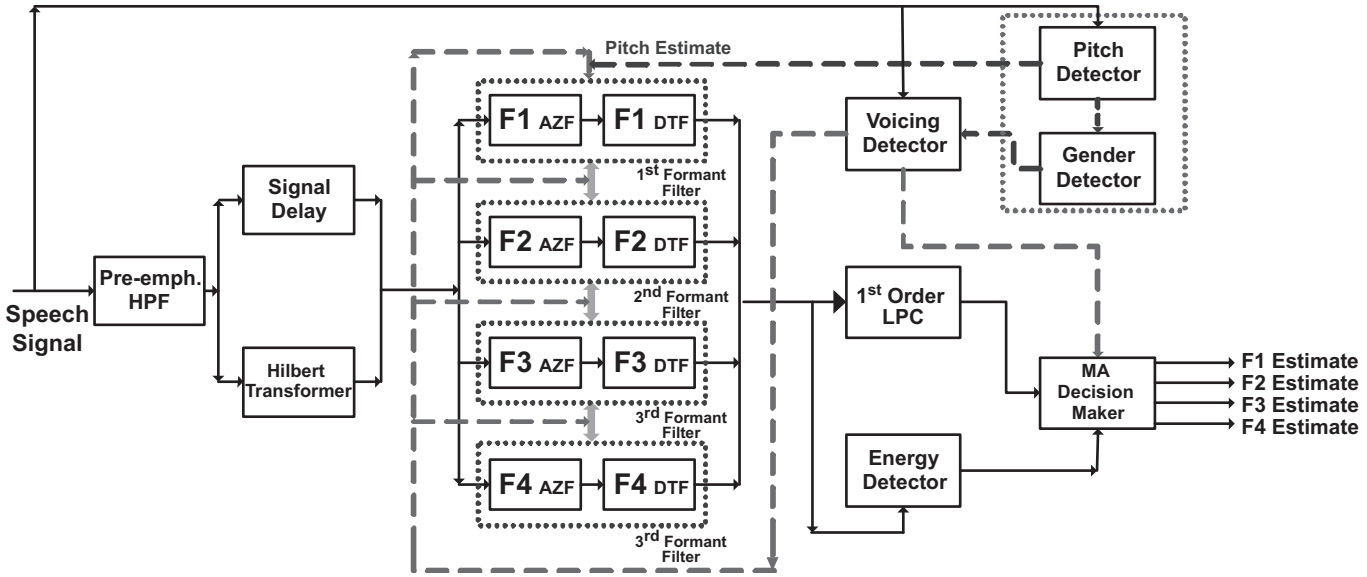


Fig. 1. Block diagram of the formant tracker. The formant tracker relies on an adaptive filterbank to separate each formant frequency region prior to spectral estimation.

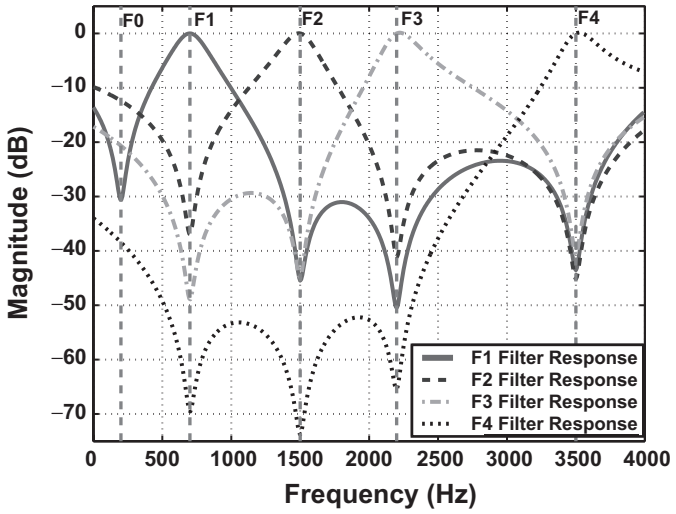


Fig. 2. The magnitude-frequency response of the four formant filters at a particular time when the pitch ($F0$) is set to 200 Hz, the first formant frequency ($F1$) is set to 700 Hz, the second formant frequency ($F2$) is set to 1500 Hz, the third formant frequency ($F3$) is set to 2200 Hz and the fourth formant frequency ($F4$) is set to 3500 Hz. Each formant filter acts as an adaptive bandpass filter, spectrally isolating the four formants.

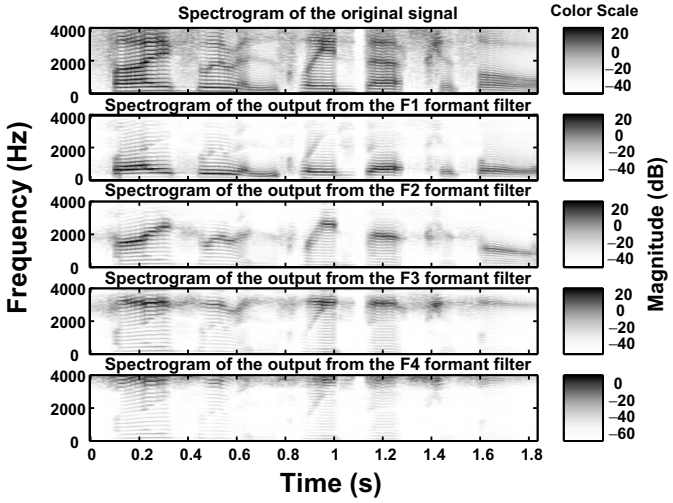


Fig. 3. Spectrograms of the original speech signal and the filtered signals from the four adaptive formant filters.

The LPCs are only calculated from the bands if the entire previous 20-ms window of the speech signal is voiced (as determined by the voicing detector described in Section II-E below) and the energy in a particular band is above the energy threshold for that band, as described in the next section.

C. Adaptive Energy Detector

After the speech signal has been filtered using the adaptive bandpass filterbank, the RMS energy of the signal over the previous 20 ms is calculated for each band. In order to estimate a particular formant frequency from the spectrum (instead of assigning the moving average value to that formant frequency), the energy calculated in that formant band has to be above an '*energy threshold level*', in addition to that speech segment being voiced. The energy threshold level for each of the formant frequencies is different and is adapted to long term changes in the spectral energy of the formant frequency bands. Gradual adjustment of the threshold levels prevents long term errors to the energy detector and allows the algorithm to recover quickly from brief loud sounds. The energy threshold level ET_{Fi} for the i^{th} formant is updated during every voiced segment of speech according to:

$$ET_{Fi}[n] = ET_{Fi}[n-1] - 0.002\left(ET_{Fi}[n-1] - E_{Fi}[n]\right), \quad (6)$$

where $ET_{Fi}[n]$ is the energy threshold level (in dB) of the i^{th} formant frequency at time index n and $E_{Fi}[n]$ is the RMS energy (in dB) of the previous 20 ms of the signal at the output

TABLE I
INITIAL FORMANT ENERGY THRESHOLD LEVELS.

Initial formant energy threshold	Level (dB)
$ET_{F1_{\text{init}}}$	-35
$ET_{F2_{\text{init}}}$	-40
$ET_{F3_{\text{init}}}$	-45
$ET_{F4_{\text{init}}}$	-50

of the i^{th} formant filter.

The initial energy threshold levels $ET_{Fi_{\text{init}}}$ for each formant Fi are set at the start of the algorithm. Various initial energy threshold levels $ET_{Fi_{\text{init}}}$ were tested and the best results were obtained using the values given in Table I.

D. Gender Detector

Gender detection is based on a simple and fast pitch estimator [27]. This algorithm uses an autocorrelation based approach with center-clipping, for pitch frequency estimation. Center-clipping the signal makes the periodicity of the speech signal more prominent while suppressing the interaction between the pitch frequency and the first formant frequency, thus increasing the accuracy of the pitch frequency estimates [27], [28].

A 50-ms segment of speech is broken up into seven 20-ms frames overlapping by 5 ms. After the signal in a particular frame has been center clipped, its autocorrelation, $R_n[m]$, is calculated and the location of the highest peak, p, of the autocorrelation function is located. If $R_n[p]$ is greater than $0.4 \times R_n[0]$, then the pitch period is computed from p. Otherwise, the segment is classified as being unvoiced and its pitch is set to 0 Hz. The range of acceptable values for the pitch frequency is between 60 and 320 Hz, and if the calculated value of the pitch for a particular frame is outside this range then the pitch estimate for that frame is set to 0 Hz. The pitch for the entire segment is obtained by median-filtering the pitch estimates from all the frames within that segment.

The gender $G[n]$ is updated every 20 ms (160 samples); the speaker is considered to be male ($G[n] = 0$) if the average pitch frequency is below 180 Hz and is set to female ($G[n] = 1$) if it is great than or equal to 180 Hz. The average pitch frequency of each segment is also used by the $F1$ formant filter for the placement of the additional zero at the pitch frequency location, as described in Section II-A above.

E. Voicing Detector

A block diagram of the voicing detector is shown in Fig. 4. The voicing detector provides the formant tracking algorithm with a reliable sample-by-sample decision on whether the preceding 20-ms speech segment is voiced or unvoiced. The low-frequency to high-frequency energy ratio serves as the primary means of determining if a speech segment is voiced or unvoiced. Functionality has been built into the voicing detector to prevent it from switching its decisions spuriously, e.g., as a result of short-term fluctuations in the speech spectrum. Parameters of the voicing detector need to be adapted to so that it functions well for both male and female speakers; the gender detector provides regular updates to the voicing detector about the gender of the speaker so that the voicing detector parameters can be updated.

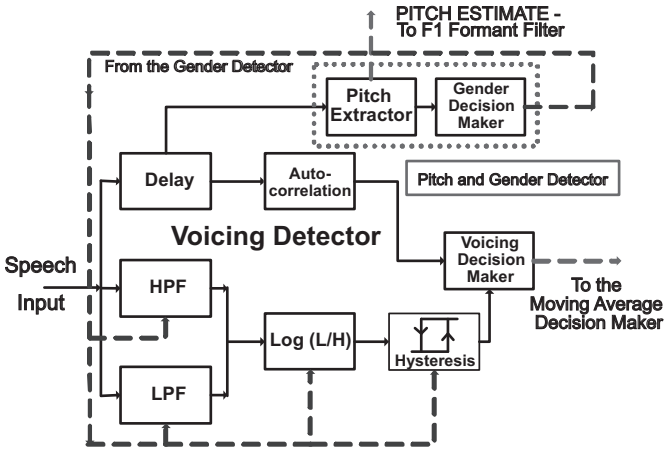


Fig. 4. Block diagram of the voicing detector designed to identify voiced segments of speech.

In the voicing detector, the original speech signal (the real-valued signal without pre-emphasis) is filtered into two different frequency bands by passing it through a highpass filter (HPF) and a lowpass filter (LPF) with the same cutoff frequency F_{vd}. After the signal is filtered into the two frequency bands, the RMS energy of the previous 20 ms of the lower and higher frequency bands is normalized by the square root of the filter bandwidths, i.e., divided by $\sqrt{F_{\text{vd}}}$ for the low-frequency band and by $\sqrt{F_s/2 - F_{\text{vd}}}$ for the high-frequency band. The log ratio of the normalized low-frequency energy to high-frequency energy is then calculated. The windowed signal segment is classified as voiced when the log ratio exceeds a threshold level.

The value of F_{vd} depends on the estimated gender of the speaker. For the large number of values tested, the best results were obtained when F_{vd} was set to 700 Hz for male speech and 1120 Hz for female speech. Every 20 ms the voicing detector obtains updates of the estimated gender of the speaker and is able to modify F_{vd} if the gender changes. The cut-off frequency F_{vd} is adapted slowly so that transient effects are limited. The algorithm is configured to shift the cut-off frequency F_{vd} from 700 to 1120 Hz (from that for a male speaker to that for a female speaker) or vice versa over ~ 44 ms according to the equation:

$$F_{\text{vd}}[n] = \begin{cases} F_{\text{vd}}[n-1] - 1.2, & \text{if } G[n] = 0 \\ & \quad \text{and } F_{\text{vd}}[n-1] > 700, \\ F_{\text{vd}}[n-1] + 1.2, & \text{if } G[n] = 1 \\ & \quad \text{and } F_{\text{vd}}[n-1] < 1120, \\ F_{\text{vd}}[n-1], & \text{otherwise,} \end{cases} \tag{7}$$

where $F_{\text{vd}}[n]$ is the cut-off frequency and $G[n]$ is the estimated gender (zero for male and one for female) at sample index n.

The log energy ratio is reliable and accurate only for phonemes with frequency components that do not vary too much over time. The presence of transient frequency components can make the voicing detector results oscillate too

quickly between the voiced and unvoiced states. To avoid these spurious oscillations between the voiced and unvoiced states, Bruce and colleagues [20] proposed a threshold with hysteresis. This allows changes in the voicing state (from voiced to unvoiced or vice versa) only if the state of the current sample changes from the previous sample and the current sample has a log ratio exceeding a set threshold level.

If the previous 20-ms window (i.e., up to sample n–1) was unvoiced and the current 20-ms window (i.e., up to sample n) has a log energy ratio greater than a set threshold level ($LT_{\rm v}$), then the current sample is assigned as being voiced, i.e., the switch from unvoiced to voiced state occurs only if the log energy ratio is greater than the proper threshold level. If the previous window was voiced and the current window has a log ratio less than a set threshold level ($LT_{\rm u}$), then the current sample is assigned as being unvoiced, i.e., the switch from voiced to unvoiced state occurs only if the log energy ratio is below the proper threshold level. These threshold levels depend on the gender of the speaker and have to be changed as the gender of the speaker changes. From the range of values tested, the best results for the voicing detector were obtained when $LT_{\rm v}$ was set to 0.1 for males and 0.2 for females and when $LT_{\rm u}$ was set to -0.2 for males and -0.3 for females. If the gender of the speaker changes, the threshold levels are updated over 40 ms (to avoid any transient effects) according to the equations:

$$LT_{\rm v}[n] = \begin{cases} LT_{\rm v}[n\text{–}1] - 3.125\times10^{-4}, & \text{if } G[n]=0 \\ & \text{and } LT_{\rm v}[n\text{–}1] > 0.1, \\ LT_{\rm v}[n\text{–}1] + 3.125\times10^{-4}, & \text{if } G[n]=1 \\ & \text{and } LT_{\rm v}[n\text{–}1] < 0.2, \\ LT_{\rm v}[n\text{–}1], & \text{otherwise,} \end{cases} \tag{8}$$

and

$$LT_{\rm u}[n] = \begin{cases} LT_{\rm u}[n\text{–}1] + 3.125\times10^{-4}, & \text{if } G[n]=0 \\ & \text{and } LT_{\rm u}[n\text{–}1] < -0.2, \\ LT_{\rm u}[n\text{–}1] - 3.125\times10^{-4}, & \text{if } G[n]=1 \\ & \text{and } LT_{\rm u}[n\text{–}1] > -0.3, \\ LT_{\rm u}[n\text{–}1], & \text{otherwise.} \end{cases} \tag{9}$$

In order to avoid erroneous voicing detection in the presence of background noise with a random lowpass spectrum over the short term, the voicing detector algorithm performs an autocorrelation-based test to determine if the energy in the lower frequency band is due to short-term colored noise or due to some other lowpass signal that may be speech. The autocorrelation of the previous 20 ms of the signal is calculated. The signal is classified as voiced only if, in addition to passing the log energy ratio test, the autocorrelation at any lag ($m \neq 0$) is greater than the autocorrelation threshold multiplier α times the autocorrelation at zero lag ($m = 0$). The value of the multiplier α is different for male and female speakers. For the range of values tested, the best results were obtained when α was set to 0.25 for female speakers and 0.6 for male speakers. If the gender of the speaker changes, the multiplier $\alpha[n]$ is updated over ~ 44 ms (to avoid any transient effects) according to the equation:

$$\alpha[n] = \begin{cases} \alpha[n-1] + 0.001, & \text{if } G[n]=0 \\ & \text{and } \alpha[n-1] < 0.6, \\ \alpha[n-1] - 0.001, & \text{if } G[n]=1 \\ & \text{and } \alpha[n-1] > 0.25, \\ \alpha[n-1], & \text{otherwise.} \end{cases} \tag{10}$$

Independent testing of the voicing detector algorithm was conducted using both synthesized sentences and recorded speech sentences from the TIMIT database. The results show that the voicing detector performs very well for both male and female speakers and has a delay of approximately 10 ms from the actual onset of voicing to the detection of voicing. The voicing detector is fairly robust and encounters very little spurious switching between the voiced and unvoiced states.

F. Moving Average Decision Maker

The moving average decision maker has two purposes:

- to calculate and update the moving average value of each formant frequency, and
- to determine whether to assign the LPC-estimated value for the current formant frequency estimate or to decay to the moving average value for each formant frequency.

The moving average decision maker assigns the estimated value to the formant frequencies (from the LPCs) only when every sample in the 20-ms LPC window (i.e., 160 samples) is determined to be voiced according to the voicing detector and the energy of the formant is above its respective threshold level (described in Section II-C above). If not all of the windowed segment is voiced or if the energy of a particular formant is below its respective threshold level, then the current value of the formant frequency decays toward the moving average value for that formant frequency according to:

$$Fi[n] = Fi[n-1] - \left(0.002\left(Fi[n-1] - Fi_{\rm MA}[n-1]\right)\right), \tag{11}$$

where $Fi[n]$ is the formant estimate the $i^{\rm th}$ formant frequency at time index n and $Fi_{\rm MA}[n-1]$ is the previous value of the moving average for the $i^{\rm th}$ formant frequency. The update rule for the moving average value of each formant frequency is:

$$Fi_{\rm MA}[n] = \frac{1}{n}\sum_{k=1}^{n} Fi[k], \tag{12}$$

where $Fi_{\rm MA}[n]$ is the moving average value for the $i^{\rm th}$ formant frequency at time n and $Fi[n]$ is the estimate of the $i^{\rm th}$ formant frequency at time n.

G. Limitations on the Proximity of Formant Frequencies

The frequency response of a formant filter becomes distorted when the poles and zeros are too close to each other. Therefore, the formant tracking algorithm limits how close the formant frequency estimates can come to each other. The algorithm does not allow $F1$ to be less than 150 Hz from the pitch frequency $F0$ at any time. Similarly, $F2$, $F3$ and $F4$ are limited from being less than 300, 400 and 500 Hz, respectively, from their lower neighbors.

TABLE II
MEANS AND STANDARD DEVIATIONS OF INITIAL FORMANT FREQUENCY VALUES FOR TESTING PURPOSES.

Initial formant estimate	μ_{Fk} (Hz)	σ_{Fk} (Hz)
$F0_{\text{init}}$	150	40
$F1_{\text{init}}$	458	154
$F2_{\text{init}}$	1535	469
$F3_{\text{init}}$	2705	468
$F3_{\text{init}}$	3819	485

III. TESTING AND RESULTS

Rigorous and systematic testing of the formant tracking algorithm was conducted in order to find best values for the operating parameters as well as to ensure that the algorithm performs well under various levels and types of background noise. The algorithm has been tested using various synthesized speech signals as well as a large number of signals from the TIMIT recorded speech database. Testing with synthesized sentences allows quantitative analysis of the performance of the formant tracker because the formant frequency values of the synthesized speech signals are known. The TIMIT database speech signals are recorded from actual speakers and therefore sound more natural than the synthesized speech signals. However, because the actual formant frequency values of the TIMIT database speech signals are unknown, only qualitative analysis of the results can be performed through visual inspection of the spectrograms.

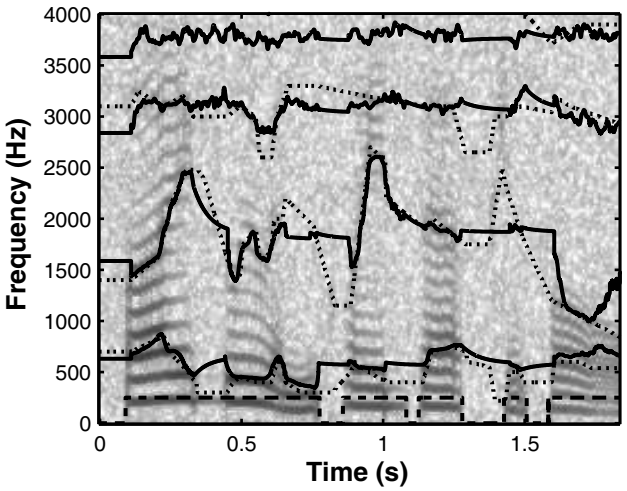
For testing purposes, the formant tracking algorithm was initialized with formant frequency estimates set to random values drawn from distributions that approximate the true distributions of naturally-occurring formant frequencies. The mean and standard deviation in frequency of each formant $F1$–$F4$ and the pitch $F0$ were calculated from formant frequency values of male and female synthesized utterances of the sentence "Five women play basketball." During testing, the initial values for the four formant frequencies and the pitch frequency were drawn from Gaussian distributions $\mathcal{N}(\mu_{Fk}, \sigma_{Fk})$, where the mean μ_{Fk} and standard deviation σ_{Fk} for each formant Fk, for $k = 1, 2, 3,$ or 4, or for the pitch $F0$, are given in Table II. The initial values were rejected and new values drawn if any value was negative or was greater than the Nyquist frequency ($F_s/2 = 4$ kHz), or if the values for the frequencies of $F0$–$F4$ were not in the correct order (i.e., ascending values).

Quantitative performance of the algorithm for the synthesized sentences was measured in terms of the root mean squared error (RMSE) in units of Hz between the actual and estimated formant frequencies. The RMSE was measured only for those time indices where the target speaker's speech was voiced and had sufficient energy for spectral estimation, and the approximate average delay of the formant trackers (10 ms) was compensated for when computing the RMSE.

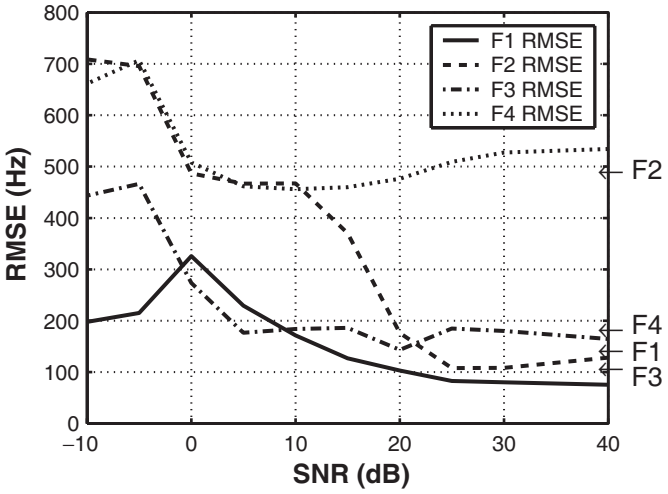
A. Testing in the Presence of White Noise

The operation of the algorithm was tested and analyzed in the presence of background additive white Gaussian noise (AWGN) at signal-to-noise ratios (SNRs) from 40 dB to -10 dB, for various synthesized and TIMIT database speech signals (for both male and female speakers).

Figure 5a) shows the spectrogram of a female synthesized speaker saying "Five women played basketball" in the presence of AWGN at 20 dB SNR. On the spectrogram, the actual formant frequencies are plotted as dotted lines, the estimated formant frequencies for an example trial are plotted as solid lines and the voicing decision is plotted as a dashed line. The speech is unvoiced when the voicing decision is 'low' (zero), and it is voiced when the voicing decision is 'high' (non-zero). Note that the approximate average delay of the formant trackers and voicing detector (10 ms) is compensated for when plotting the formant estimates against the spectrograms and true formant trajectories.



a)



b)

Fig. 5. Formant tracking results for a synthesized female speaker saying "Five women played basketball" in the presence of background white Gaussian noise. a) Spectrogram, estimated (solid lines) and actual (dotted lines) formant frequencies, and voicing decision (dashed line) at an SNR of 20 dB. b) Mean RMSE (in Hz) over 25 trials as a function of SNR (in dB). The arrows indicate the standard deviations of the actual formant frequency values (in Hz).

It can be seen that during the voiced segments of speech the algorithm performs well in tracking the actual formant

frequencies, including fairly rapid formant transitions. During unvoiced speech segments the algorithm decays the formant tracks towards the moving average values of the formant frequency estimates, rather than attempting spectral estimation. At the onset of a new voiced segment of speech, the formant tracker quickly re-acquires the correct formant trajectory.

The algorithm was tested using different male and female speech sentences in various SNRs, in the presence of background AWGN. In general, the formant frequencies were estimated accurately and the algorithm was relatively robust. Figure 5b) shows the RMS error between the actual and the estimated formant frequencies for the same sentence as in Fig. 5a) in the presence AWGN at various SNRs. 25 trials were conducted, with a different AWGN sequence and different initial formant estimates each time, and the figure shows the mean RMSEs for the 25 trials.

The arrows on the right hand side of Fig. 5b) indicate the standard deviations of the actual formant frequencies (calculated using those points for which the RMSE of the estimated formant frequencies is calculated). At high SNRs, the RMSEs for $F1$ and $F3$ are close to the standard deviations of the actual formant frequencies, and the RMSE for $F2$ is much less than the standard deviation of actual $F2$ frequencies. The RMSE for $F4$ is much higher than the standard deviation of actual $F4$ frequencies at all SNRs because the actual $F4$ frequency often exceeds the Nyquist frequency ($F_s/2 = 4$ kHz) in this sentence and consequently the formant cannot be tracked during these times. At 0 dB SNR the RMSE for $F1$ grows to around two times the standard deviation of actual $F1$ frequencies. Below 0 dB SNR, the $F1$ RMSE drops to around the standard deviation for $F1$ because at such low SNRs the algorithm decays to the moving average values of the formant frequencies instead of spectrally estimating them via linear prediction. The RMSE for $F2$ increases to around the standard deviation of actual $F2$ frequencies at SNRs between 0 and 10 dB and exceeds the standard deviation somewhat at lower SNRs. This validates the use of the moving average decision maker when spectral estimates are not reliable. Another justification for decaying to moving average values during unvoiced speech is that it ensures that the formant frequency estimates vary smoothly as the speech changes from voiced to unvoiced and vice versa, even at low SNRs.

Similar performance was found for a *male* synthesized speaker in AWGN, except that the mean RMSEs for $F4$ were substantially lower than for the female speaker, because the true $F4$ frequencies for the male speaker remain below the Nyquist frequency and can consequently be tracked with some reliability [18].

Despite the good mean RMSE results, there were some instances where the random initial formant frequencies values (drawn from the distributions described above) were far enough from the formant frequencies of the first phoneme in the test sentence that the $F2$ and $F3$ trackers had difficulty in finding the correct formant tracks. Table III shows the number of trials out of 25 for which the RMSE for $F2$ was greater than 600 Hz, indicating that $F2$ and $F3$ were poorly tracked in that trial. At high SNRs this occurred only very infrequently (zero or one times out of 25 trials). Poorly-tracked trials were more prevalent at low SNRs, particularly for the female test sentence in which (i) the formants frequencies are more widely spaced and (ii) $F4$ is often above the Nyquist frequency.

TABLE III
NUMBER OF POORLY-TRACKED TRIALS OUT OF 25 FOR THE FEMALE AND MALE VERSIONS OF "FIVE WOMEN PLAYED BASKETBALL" IN THE PRESENCE OF BACKGROUND AWGN.

SNR (dB)	-10	-5	0	5	10	15	20	25	30	40
Female	16	17	0	2	0	0	0	0	0	1
Male	9	5	6	5	1	0	0	0	1	0

B. Testing in the Presence of a Single Background Speaker

The performance of the algorithm was evaluated in the presence of a single male or female background (competing) speaker at SNRs from 40 dB to -5 dB. This scenario is challenging for the algorithm, because over a particular short time period the background speaker may contribute significant energy to the formant frequency regions of the target speaker, especially at lower SNRs. This may cause the algorithm to start tracking the formant frequencies of the background speaker instead of those of the primary speaker. Furthermore, at very low SNRs (0 dB and below), the formant tracking algorithm may start exclusively tracking the formant frequencies of the background speaker because the energy from the background speaker will be greater than that from the target speaker.

Figure 6a) shows the spectrogram and estimated and actual formant frequencies of a synthesized male speaker saying "Five women played basketball" in the presence of a competing single female recorded (TIMIT) background speaker at an SNR of 5 dB. At such a low SNR with a single competing background speaker, the voicing detector determines that signal is voiced for almost the entire duration of the signal. During the silent and unvoiced segments of the target sentence, the algorithm tracks the formants of the background female speaker and then returns to accurately tracking the target formants in the strongly voiced phonemes of the synthesized sentence. An example of this is seen in Fig. 6a) in the period $\sim$ 0.6–0.7 s, where the $F1$ and $F2$ estimates rise up in frequency to the formants of the background female speaker and then return back down to the formants of the synthesized male speaker.

Figure 6b) shows the mean RMSEs for 25 trials of the same synthesized male speaker sentence in the presence of a single female background speaker from the TIMIT database saying "How do we define it?" at various SNRs. The background sentence was identical in each of the 25 trials, but the initial formant estimates were different for each trial. Similar to the previous example, at high SNRs the mean RMSEs for $F1$, $F2$ and $F4$ are typically around the standard deviations of their respective actual formant frequencies, and as the SNR falls to 0 dB the RMSEs increase above the standard deviation values. The mean RMSEs for $F2$ are substantially below the actual $F2$ standard deviation at most SNRs. At 0 dB SNR the performance of the algorithm degrades and the RMSE for $F2$ rise above the standard deviation.

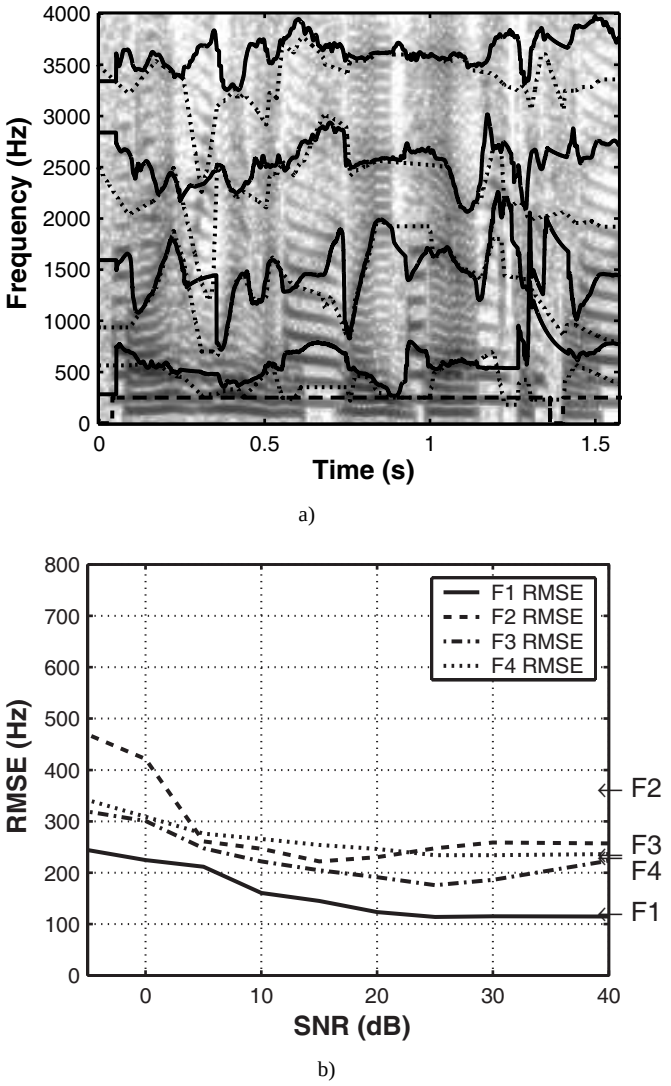


Fig. 6. Formant tracking results for a synthesized male speaker saying "Five women played basketball" in the presence of a single recorded (TIMIT) female speaker. a) Spectrogram, estimated (solid lines) and actual (dotted lines) formant frequencies, and voicing decision (dashed line)at an SNR of 5 dB. b) Mean RMSE (in Hz) over 25 trials as a function of SNR (in dB). The arrows indicate the standard deviation of the actual formant frequency values (in Hz).

C. Testing in the Presence of Multiple Background Speakers

The performance of the algorithm was also tested in the presence of multiple background competing speakers (background babble) at SNRs from 40 dB to -5 dB. This noise source has characteristics somewhere between the previous two test cases: the multiplicity of speakers produces a flatter short-term spectrum and less temporal modulation than a single speaker, but it has greater spectral and temporal modulation than AWGN.

Figure 7 shows mean RMSEs for 25 trials of a synthesized male speaker saying "Once upon a midnight dreary, while I pondered weak and weary, over many a…" in the presence of background babble at various SNRs. The background babble was identical in each of the 25 trials, but the initial formant estimates were different for each trial. At high SNRs, the RMSEs for $F2$ are below the standard deviation of the actual formant frequencies. The mean RMSE for $F2$ grows to equal the standard deviation as the SNR drops to 0 dB, but it falls slightly at -5 dB to just below the standard deviation.

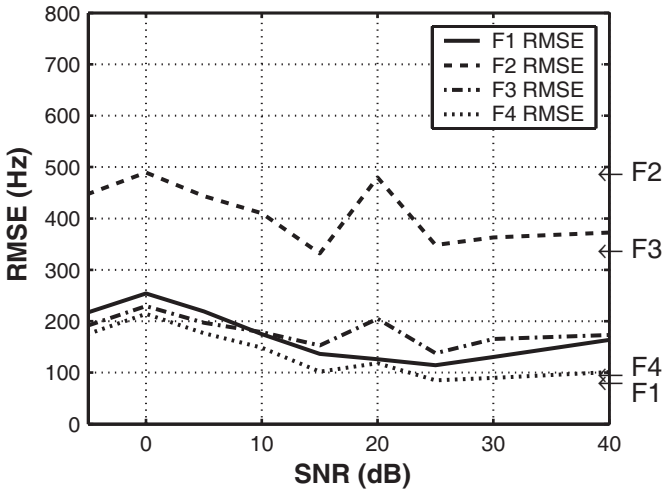


Fig. 7. Mean RMSE (in Hz) over 25 trials as a function of SNR (in dB) for a synthesized male speaker saying "Once upon a midnight dreary, while I pondered weak and weary, over many a…" in the presence of background babble. The arrows indicate the standard deviation of the actual formant frequency values (in Hz).

D. Testing for Speaker Variability

One of the main features of the algorithm is its ability to track formant frequencies for both male and female speakers. The results for the previous test cases indicate that the algorithm works quite well for both genders. However, in this test case the response of the algorithm to a change in the gender of the speaker was evaluated, to see if it can seamlessly switch between tracking formant frequencies for either gender.

Figure 8 shows a spectrogram for the transition between two concatenated TIMIT database sentences in the presence of background white noise at an SNR of 25 dB. The start of the signal is a male speaker saying "Gus saw pine trees and redwoods on his walk through Sequoia national forest" followed by a female speaker saying "Don't ask me to carry an oily rag like that." The switch from the male speaker to the female speaker occurs at approximately $t = 4.1$ s. On the spectrogram, the estimated formant frequencies are plotted as solid lines and the gender decision is plotted as a dashed line. The gender detector correctly estimates that the speaker is male where the dashed line is low (before $t \approx 4.1$ s) and that the speaker is female where the dashed line is high (after $t \approx 4.1$ s). The algorithm performs well during the transition from male to female speaker and continues to provide smooth and accurate formant frequency estimates as it starts tracking the formants of the female speaker.

E. Other Tests

The algorithm was tested using both synthesized and TIMIT database recorded speech signals under various other types and levels of background noise conditions, including:

- background music,

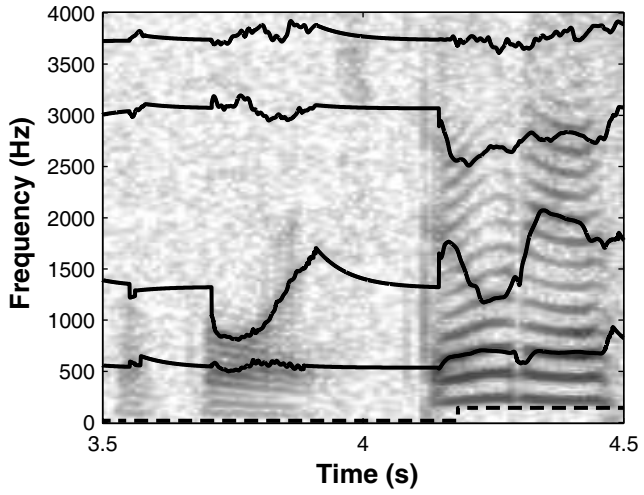


Fig. 8. Spectrogram, formant frequency estimates (solid lines) and gender decision (dashed line) for a transition between a male and a female speaker (at $t \approx 4.1$ s) in the presence of AWGN at an SNR of 25 dB. Where the dashed line is low the gender detector has determined that the speaker is male, and where the dashed line is high it has decided that a female has begun speaking.

- background environmental sounds,
- background traffic noise,
- fading of the target speech signal, and
- reverberant acoustic environments.

Further details of the testing and results can be obtained in [18].

IV. Discussion and Conclusions

As described in the Introduction, in order to use this algorithm for CEFS amplification in hearing aids the formant frequency estimates have to be smooth and accurate and must be computed with relatively little time delay. Detailed analysis has shown that the algorithm provides fairly accurate formant frequency estimates at moderate to high SNRs and is robust to real-life noise conditions such as additive white Gaussian noise, a single background speaker (of the same or different gender), multiple background speakers, reverberant acoustic environments, etc. The algorithm provides mostly smooth formant frequency estimates and recovers quickly after erroneous estimates to return to tracking actual formant frequencies in the speech signal. Furthermore, the algorithm has been designed to operate in real-time and estimate formant frequencies from continuous speech for both male and female speakers. Therefore, it can be concluded that the formant tracking algorithm presented in this paper is suitable for CEFS amplification.

There have been some problems identified with the formant tracker. It was observed that occasionally the $F2$ and $F3$ trackers had difficulty finding the correct formant tracks if the initial formant estimate values were far from the actual formant frequencies. In continuous operation of the algorithm, this might occur sometimes when there is a switch in the gender of the speaker or if the formant tracker is perturbed by transient background noise while the voicing detector reports that the signal is voiced. One solution to this problem might be to place limits on range of values that formant frequencies estimates can take, depending on the current gender estimate from the gender detector. Additionally, the algorithm occasionally gives 'choppy' and oscillating formant frequency estimates. This is an undesirable result because the actual formant frequencies of speech normally vary slowly with time and have smooth transitions. This problem is normally only encountered when the SNR is very low (typically below 5 dB) and occurs due to the algorithm tracking the excess energy added outside the formant frequency regions from the background noise source. Possible solutions to this problem may be to smooth the estimates or to incorporate additional logical limitations to prevent abnormal jumps from occurring in formant frequency estimates in the first place. Another improvement may be to modify the formant filters to have variable bandwidths that are dependent on the magnitudes of the poles estimated by the linear prediction analysis. This may further improve the formant frequency estimates during rapid formant transitions at high SNRs, but the performance of the algorithm at low SNRs would likely remain unchanged. Despite these limitations, the overall performance of the algorithm is better than those of traditional formant estimation techniques [18].

Recently, substantial improvements have been made over traditional formant tracking methods. Several approaches have incorporated more sophisticated modeling of vocal tract resonances than conventional linear prediction [29]–[33]. Another technique uses the spectral differential phase spectrum rather than the Fourier spectrum [34]. In contrast to simple logical peaking picking, many new algorithms implement estimation and tracking techniques such as concurrent curve formation [35], probabilistic estimation techniques (such as the estimation-maximization algorithm) [29], [36], [37], 1D and 2D hidden Markov models (HMMs) [38]–[42], Kalman filtering [31], [41] and particle filtering [33]. However, the computational complexity and the signal delay for most of these techniques greatly exceed those of the algorithm presented in this paper. Furthermore, it remains for most of these methods to be tested for robustness in background noise. One exception is [41], in which the combination of LP-spectral subtraction and Kalman filtering was found to produce much more robust estimation of formants in background car and train noise than a 2D HMM algorithm.

Although it was found in [18] that formant tracking based on a highly simplified model of the auditory periphery [17] was not robust to background noise, a more physiologically accurate model has previously been shown to have a robust representation of formants in white noise [43]. It would be of interest to see if the newer estimation and tracking methods listed above could be applied to the output of such a physiological model to produce robust formant estimates for a variety of background noises. However, once again the computational requirements and the signal delay are likely to exceed those of the formant tracker presented in this paper.

Although the algorithm developed in this paper is primarily designed to meet the criteria for CEFS amplification, other applications for it do exist, such as automatic speech recognition [29], speech synthesis [44], speaker normalization for automatic speech recognition [45], voice conversion [40],

speaker identification [46] and speech coding [47]. Some of these applications utilize phonemic segmentation before formant estimation, in which case the voicing detector and moving average decision maker could be removed from our system, and the remaining algorithm could be applied just to segments that are identified as voiced phonemes.

ACKNOWLEDGMENT

The authors would like to thank Siddharth Das for assistance in development and testing of the voicing detector section of the algorithm.

REFERENCES

[1] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.*, vol. 24, no. 2, pp. 175–184, Mar. 1952.

[2] L. C. Pols, L. J. van der Kamp, and R. Plomp, "Perceptual and physical space of vowel sounds," *J. Acoust. Soc. Am.*, vol. 46, no. 2, pp. 458–467, Aug. 1969.

[3] P. F. Assmann, "The role of formant transitions in the perception of concurrent vowels," *J. Acoust. Soc. Am.*, vol. 97, no. 1, pp. 575–584, Jan. 1995.

[4] ——, "Modeling the perception of concurrent vowels: Role of formant transitions," *J. Acoust. Soc. Am.*, vol. 100, no. 2 Pt 1, pp. 1141–1152, Aug. 1996.

[5] R. N. Ohde, "The development of the perception of cues to the [m]-[n] distinction in CV syllables," *J. Acoust. Soc. Am.*, vol. 96, no. 2 Pt 1, pp. 675–686, Aug. 1994.

[6] A. K. Nábělek, Z. Czyzewski, and H. Crowley, "Cues for perception of the diphthong /aɪ/ in either noise or reverberation. Part I. duration of the transition," *J. Acoust. Soc. Am.*, vol. 95, no. 5 Pt 1, pp. 2681–2693, May 1994.

[7] A. M. Liberman, P. C. Delattre, F. S. Cooper, and L. J. Gerstman, "The role of consonant-vowel transitions in the perception of the stop and nasal consonants," *Psychological Monographs*, vol. 68, pp. 1–13, 1954.

[8] R. L. Miller, J. R. Schilling, K. R. Franck, and E. D. Young, "Effects of acoustic trauma on the representation of the vowel /ɛ/ in cat auditory nerve fibers," *J. Acoust. Soc. Am.*, vol. 101, no. 6, pp. 3602–3616, June 1997.

[9] M. B. Sachs, I. C. Bruce, R. L. Miller, and E. D. Young, "Biological basis of hearing-aid design," *Ann. Biomed. Eng.*, vol. 30, no. 2, pp. 157–168, Feb. 2002.

[10] J. R. Schilling, R. L. Miller, M. B. Sachs, and E. D. Young, "Frequency-shaped amplification changes the neural representation of speech with noise-induced hearing loss," *Hear. Res.*, vol. 117, pp. 57–70, 1998.

[11] R. L. Miller, B. M. Calhoun, and E. D. Young, "Contrast enhancement improves the representation of /ɛ/-like vowels in the hearing-impaired auditory nerve," *J. Acoust. Soc. Am.*, vol. 106, no. 5, pp. 2693–2708, 1999.

[12] I. C. Bruce, "Physiological assessment of contrast-enhancing frequency shaping and multiband compression in hearing aids," *Physiol. Meas.*, vol. 25, no. 4, pp. 945–956, Aug 2004.

[13] J. L. Flanagan, "Automatic extraction of formant frequencies from continuous speech," *J. Acoust. Soc. Am.*, vol. 28, pp. 110–118, 1956.

[14] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Am.*, vol. 47, no. 2, pp. 634–648, Feb. 1970.

[15] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am.*, vol. 50, no. 2, pp. 637–655, Aug. 1971.

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

[17] S. W. Metz, J. A. Heinen, R. J. Niederjohn, and T. V. Sreenivas, "Auditory modeling applied to formant tracking of noise-corrupted speech," in *Proc. Intl. Conf. Industrial Electronics, Control and Instrumentation, vol. 3*, 1991, pp. 2120–2124.

[18] K. Mustafa, "Robust formant tracking for continuous speech with speaker variability," Master's thesis, McMaster University, 2003.

[19] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 240–254, May 2000.

[20] I. C. Bruce, N. V. Karkhanis, E. D. Young, and M. B. Sachs, "Robust formant tracking in noise," in *Proc. ICASSP 2002, Vol. I*. Piscataway, NJ: IEEE, 2002, pp. 281–284.

[21] K. Mustafa and I. C. Bruce, "Robust formant tracking for continuous speech with speaker variability," in *Proceedings of the Seventh International Symposium on Signal Processing and Its Applications (ISSPA), Vol. 2*. Piscataway, NJ: IEEE, 2003, pp. 623–624.

[22] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. New York: Macmillan, 1993.

[23] L. R. Rabiner and R. W. Schafer, "On the behavior of minimax FIR digital Hilbert transformers," *The Bell System Technical Journal*, vol. 53, no. 2, pp. 363–390, 1974.

[24] J. Picone, D. P. Prezas, W. T. Hartwell, and J. L. Locicero, "Spectrum estimation using an analytic signal representation," *Sig. Proc.*, vol. 15, no. 2, pp. 169–182, September 1988.

[25] T. F. Quatieri, *Discrete-Time Speech Signal Processing*. Upper Saddle River, NJ: Prentice Hall PTR, 2002.

[26] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[27] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice Hall, 1978.

[28] M. M. Sondhi, "New methods of pitch extraction," *IEEE Trans. Audio and Electroacoustics*, vol. AU-16, no. 2, pp. 262–266, 1968.

[29] L. Deng and J. Ma, "Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics," *J. Acoust. Soc. Am.*, vol. 108, no. 6, pp. 3036–3048, 2000.

[30] L. Deng, I. Bazzi, and A. Acero, "Tracking vocal tract resonances using an analytical nonlinear predictor and a target-guided temporal constraint," in *Proc. EUROSPEECH*, 2003, pp. 73–76.

[31] L. Deng, L. Lee, H. Attias, and A. Acero, "A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances," in *Proc. ICASSP 2004*, vol. 1, 2004, pp. 557–560.

[32] L. Deng, D. Yu, and A. Acero, "A quantitative model for formant dynamics and contextually assimilated reduction in fluent speech," in *Proc. InterSpeech-ICSLP*, 2004, pp. 981–984.

[33] Y. Zheng and M. Hasegawa-Johnson, "Formant tracking by mixture state particle filter," in *Proc. ICASSP 2004*, vol. 1, 2004, pp. 565–568.

[34] B. Bozkurt, T. Dutoit, B. Doval, and C. D'Alessandro, "Improved differential phase spectrum processing for formant tracking," in *Proc. InterSpeech-ICSLP*, 2004, pp. 2421–2424.

[35] Y. Laprie, "A concurrent curve strategy for formant tracking," in *Proc. InterSpeech-ICSLP*, 2004, pp. 2405–2408.

[36] P. N. Garner and W. J. Holmes, "On the robust incorporation of formant features into hidden Markov models for automatic speech recognition," in *Proc. ICASSP 1998*, vol. 1, 1998, pp. 1–4.

[37] I. Bazzi, A. Acero, and L. Deng, "An expectation maximization approach for formant tracking using a parameter-free non-linear predictor," in *Proc. ICASSP 2003*, vol. 1, 2003, pp. 464–467.

[38] A. Acero, "Formant analysis and synthesis using hidden markov models," in *Proc. EUROSPEECH*, 1999, pp. 1047–1050.

[39] K. Weber, F. de Wet, B. Cranen, L. Boves, S. Bengio, and H. Bourlard, "Evaluation of formant-like features for ASR," in *Proc. InterSpeech-ICSLP*, 2002, pp. 2101–2104.

[40] E. Turajlic, D. Rentzos, S. Vaseghi, and C.-H. Ho, "Evaluation of methods for parameteric formant transformation in voice conversion," in *Proc. ICASSP 2003*, vol. 1, 2003, pp. 724–727.

[41] Q. Yan, E. Zavarehei, S. Vaseghi, and D. Rentzos, "A formant tracking LP model for speech processing," in *Proc. InterSpeech-ICSLP*, 2004, pp. 2409–2412.

[42] F. de Wet, K. Weber, L. Boves, B. Cranen, S. Bengio, and H. Bourlard, "Evaluation of formant-like features on an automatic vowel classification task," *J. Acoust. Soc. Am.*, vol. 116, no. 3, pp. 1781–1792, 2004.

[43] L. Deng and C. D. Geisler, "A composite auditory model for processing speech sounds," *J. Acoust. Soc. Am.*, vol. 82, no. 6, pp. 2001–2012, Dec. 1987.

[44] W. Ding and N. Campbell, "Optimising unit selection with voice source and formants in the CHATR speech synthesis system," in *Proc. EUROSPEECH*, 1997, pp. 537–540.

[45] M. Lincoln, S. Cox, and S. Ringland, "A fast method of speaker normalisation using formant estimation," in *Proc. EUROSPEECH*, 1997, pp. 2095–2098.

[46] A. Park and T. J. Hazen, "ASR dependent techniques for speaker identification," in *Proc. InterSpeech-ICSLP*, 2002, pp. 478–479.

[47] J. Högberg, "Data driven formant synthesis," in *Proc. EUROSPEECH*, 1997, pp. 565–568.

Kamran Mustafa (S'02) was born in Karachi, Pakistan, in 1979. He completed his B.Sc. degree (computer engineering) from the University of New Brunswick (Fredericton, Canada) in 2001 and M.A.Sc. degree (electrical and computer engineering) from McMaster University (Hamilton, Canada) in 2003.

From 1999 to 2001 he worked at Nortel Networks in Ottawa, Canada, in various capacities within the optical networks division. He was a Research Associate in the Auditory Engineering Laboratory at McMaster University at the beginning of 2004 and is now a Systems Design Engineer with the Wireless Mesh Networks division of Nortel Networks in Ottawa, Canada. His research interests are in the areas of speech processing, hearing aids and speaker recognition.

Ian C. Bruce (S'96–M'98) was born in Bendigo, Australia, in 1969. He completed the B.E. (electrical and electronic) and Ph.D. degrees at The University of Melbourne, Australia, in 1991 and 1998, respectively.

From 1993 to 1994 he was a Research and Teaching Assistant in the Department of Bioelectricity and Magnetism, Vienna University of Technology, Austria. He was a Postdoctoral Research Fellow in the Department of Biomedical Engineering at Johns Hopkins University, Baltimore, USA, from 1998 to 2001. Presently he is an Assistant Professor in Electrical and Computer Engineering and the Barber-Gennum Chair in Information Technology at McMaster University, Hamilton, Canada. His research interests are in the areas of neural modeling, speech processing, hearing aids and cochlear implants, neurophysiology, psychophysics, and stochastic and nonlinear systems.

Dr. Bruce is a member of the Association for Research in Otolaryngology and the Acoustical Society of America.