ROBUST FORMANT TRACKING IN NOISE

Ian C. Bruce, Neel V. Karkhanis†, Eric D. Young, and Murray B. Sachs*

Center for Hearing Sciences and Department of Biomedical Engineering,
Johns Hopkins University School of Medicine, Baltimore MD 21205, USA.

ABSTRACT

While many algorithms exist for accurate extraction of formant frequencies from a speech waveform, these algorithms are not typically shown to be robust in the presence of highly-transient background noise such as competing speech waveforms. Preliminary results are presented from an algorithm using time-varying adaptive filters that appears to be robust in the presence of white, Gaussian noise or a single competing speaker over a large range of signal-to-noise ratios (quiet to -6 dB). Use of a synthesized sentence, for which the actual formant frequencies are known, permits quantitative assessment of the algorithm's accuracy as a function of signal-to-noise ratio.

1. INTRODUCTION

Signal processing methods for tracking formants typically utilize some form of spectral analysis and estimate the formant frequencies from the spectral peaks [1, 2, 3]. Such "peak picking" is made difficult when there is transient background noise at a similar amplitude to a formant or when neighboring formants are close together in frequency such that their spectral shapes overlap. Rao and Kumaresan suggested the use of adaptive filters to preprocess a speech stimulus before spectral estimation [4], such that only one formant is tracked by each adaptive filter and subsequent spectral estimator—energy from neighboring formants is ignored. We hypothesize that this algorithm would similarly be effective in reducing the influence of background noise on formant frequency estimates.

Promising results were obtained by Rao and Kumaresan for test speech stimuli in quiet, but the algorithm was not implemented in a fashion suitable for real-time application. For many applications, such as dynamic formant enhancement in a digital hearing aid [5], any algorithm must be suitable for implementation in real-time with minimal time lag in the formant estimates. Additionally, the algorithm must be robust, so that if it momentarily loses track of a formant, it can quickly refind the formant in the next segment of voiced speech.

2. IMPROVEMENTS TO THE ALGORITHM OF RAO AND KUMARESAN

Initial testing of a simulated real-time version of the Rao and Kumaresan algorithm showed that the formant trackers could be misled by unvoiced segments of the sentence, where there are no formants, or by background noise or silence when the energy in an individual formant was negligible. We propose adding a voicing detector and a formant energy detector to test for each of these cases.

Figure 1 shows our design for an individual formant tracker. The formant trackers are all identical, except that the initial formant frequencies estimates are distributed across the range of speech frequencies, such that each tracker should be able to find an individual formant at the onset of voiced speech. The acoustic signal is first pre-emphasized by a high-pass filter (HPF) to equalize roughly the energy in each formant. An approximate analytic version of the signal is then produced to improve the accuracy of the spectral estimation performed later in the algorithm [6]. The signal is subsequently passed through an all-zero filter (AZF), in which the zeros are set to the latest estimates of the *other* formant tracker estimates, suppressing their influence in this formant tracker. The AZF is followed by a single-pole dynamic tracking filter (DTF), in which the pole frequency is set to the latest frequency estimate from the same formant tracker, emphasizing signal energy near that frequency. A first-order linear prediction coding (LPC) analysis is performed to obtain the current formant frequency estimate. Because formant transitions are relatively slow, adaptive filtering (AZF + DTF) before the LPC analysis helps the tracker to continuously follow a single formant and not be distracted by the other formants. The filtering should similarly assist in the presence of background noise. To improve the robustness of the algorithm, we propose taking a moving average (MA) of the formant estimates over time; when no

*Now with the Department of Electrical and Computer Engineering, McMaster University, 1280 Main Street West, Hamilton, Ontario, Canada L8S 4K1. Email: ibruce@ieee.org

†Now with the NIH Transducer Resource Center, Department of Bioengineering, Pennsylvania State University, State College PA 16802, USA.

This research was supported by NIDCD grants DC00109 and DC00023.

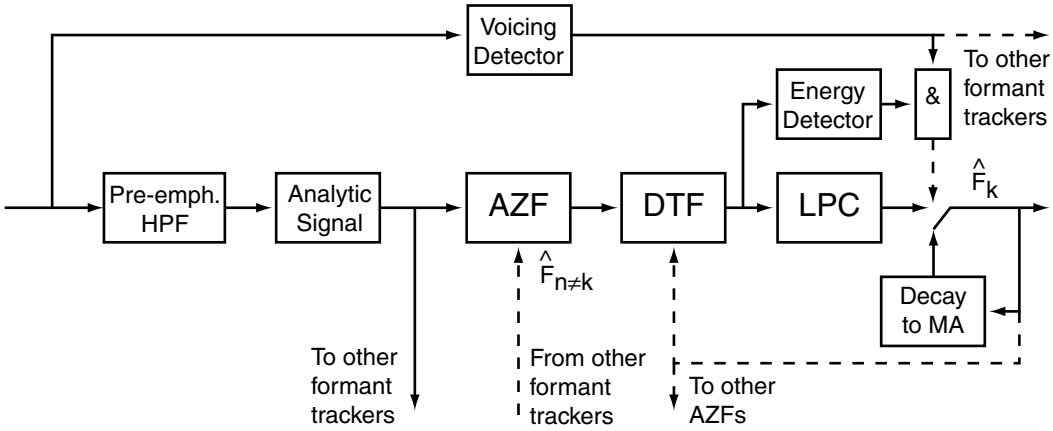

Fig. 1. Formant tracker modified from [4]. Abbreviations: high-pass filter (HPF); all-zero filter (AZF); dynamic tracking filter (DTF); linear prediction coding (LPC); moving average (MA); formant frequency estimate $\left(\hat{F}_k\right)$. See the text for more details.

voiced speech is detected or the energy in a formant drops below a certain threshold, the formant tracker's frequency estimate decays back towards the moving average.

3. RESULTS

The performance of the improved algorithm for a synthesized sentence "Five women played basketball" (courtesy of R. McGowan of Sensimetrics Corp, Somerville, MA) is illustrated in Fig. 2. Formant tracker estimates are plotted in the top panel over the spectrogram of the sentence. During voiced speech (indicated by harmonic structure in the spectrogram) the algorithm quickly finds the formant frequencies and tracks any transitions in the formant frequencies. During unvoiced speech and silence the trackers decay back to the moving average of each tracker's estimates, ready for the next segment of voiced speech. The bottom panel illustrates the gain-frequency response of the combined AZF and DTF for each formant tracker at one instant in time. Each filter combination has a peak at the latest estimate of the formant frequency and zeros at the frequency estimates of the other formants.

Figure 3 shows the performance of the tracking algorithm in background white, Gaussian noise with a signal-to-noise ratio (SNR) of 6 dB. For easy visual evaluation, the same formant estimates for the sentence with background noise are plotted over the clean speech spectrogram (top panel) and over the speech plus noise spectrogram (bottom panel). The formant trackers again do an excellent job, except where the formant energy is swamped by the background noise (e.g., F2 at ~ 0.35 s and F2–F4 at ~ 1.45 s), but these instances do not prevent the trackers from finding the formants in the next segment of voiced speech.

Use of a synthesized sentence with known formant tra-

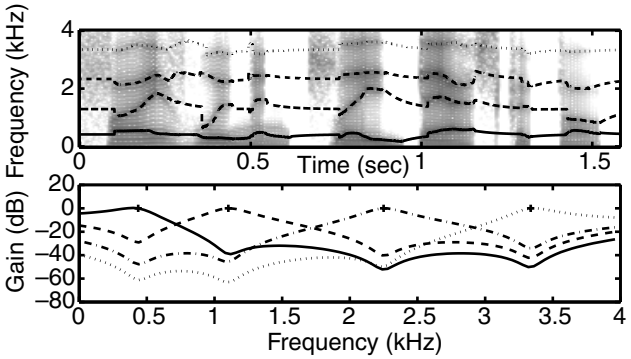

Fig. 2. Formant tracker performance in quiet for the synthesized sentence "Five women played basketball." Formant tracker estimates are plotted in the top panel over the spectrogram of the sentence. The known time delay of the formant trackers (10 ms) is compensated for, to aid comparison with the spectrogram. Combined AZF and DTF gain frequency responses are plotted (bottom panel) for each formant tracker at one instant in time. F1 (solid line), F2 (dashed line), F3 (dot-dashed line), F4 (dotted line).

jectories permits evaluation by calculating the root-mean-square error (RMSE) between the formant estimates and the known formant values from the speech synthesizer during voiced segments of speech when there is non-negligible formant energy. The known time delay of the formant trackers (10 ms) is compensated for. Plotted in Fig. 4 are RMSEs for each formant tracker as a function of SNR with white, Gaussian noise. The RMSEs increase systematically as a function of formant frequency because of the diminishing formant energy with formant frequency. The RMSEs for F1 and F2 increase fairly steadily with SNR; the F1 tracker is still fairly accurate at -6 dB SNR, while the F2 tracker

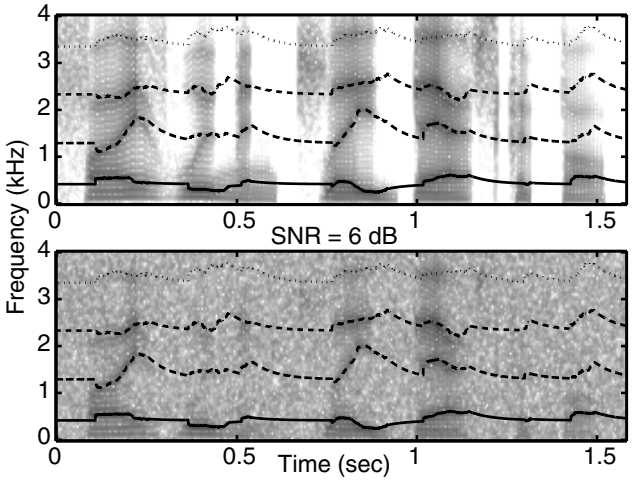

Fig. 3. Formant tracker performance in background white, Gaussian noise (at 6 dB SNR) for the synthesized sentence "Five women played basketball." Formant tracker estimates are plotted in the top panel over the spectrogram of the sentence and in the bottom panel over the spectrogram of the sentence plus background noise. The plotting convention is the same as for Fig. 2.

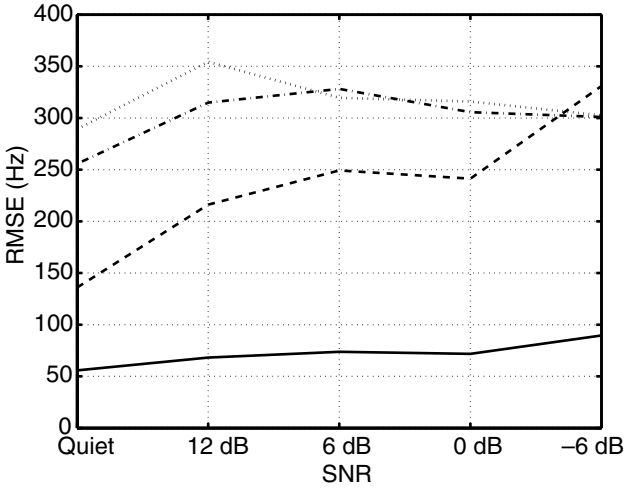

Fig. 4. RMS errors of formant trackers in background white, Gaussian noise as a function of SNR for the synthesized sentence "Five women played basketball." F1 (solid line), F2 (dashed line), F3 (dot-dashed line), F4 (dotted line).

performance has deteriorated considerably. The RMSEs for F3 and F4 increase at an SNR of 12 dB compared to the quiet condition but then decrease again at lower SNRs (6 to −6 dB). At moderate SNRs, the signal is often detected as being voiced, but the F3 and F4 trackers are distracted by the background noise. At lower SNRs, the voicing detector rarely detects the voiced segments of speech, and consequently the F3 and F4 formant trackers tend to stay at their moving averages and are less likely to wander away from the true formant values than at the moderate SNRs.

Shown in Fig. 5 is the performance of the formant trackers for the same synthesized sentence in the presence of a competing single speaker ("Don't ask me to carry an oily rag like that," from the TIMIT speech database). Identical formant tracker estimates for the combined sentences are plotted over the spectrogram of the synthesized sentence in the top panel, the competing sentence in the middle panel, and the two sentences combined at an SNR of 6 dB in the bottom panel. This SNR is sufficient for the trackers to ignore formants in the competing sentence, except when there is no formant energy in the "target" sentence; the algorithm quickly finds the target formants at the next onset of voiced speech in the target sentence.

Plotted in Fig. 6 are RMSEs as a function of SNR with the single competing speaker. The results are very similar to those for white, Gaussian noise (see Fig. 4). The RMSEs increase systematically as a function of formant frequency, and the RMSEs for F1 and F2 increase fairly steadily with SNR; the F1 tracker is still fairly accurate at −6 dB SNR, while the F2 tracker performance has deteriorated considerably. The RMSEs for F3 and F4 increase at an SNR of 12 dB but then decrease again at lower SNRs (6 to −6 dB). At moderate SNRs, the signal is often detected as being voiced because of the presence of the competing sentence, but the F3 and F4 trackers are distracted by the unvoiced high-frequency energy in the target sentence. At lower SNRs, the F3 and F4 formant trackers tend to alternate between following energy in the target and competing sentences and are less likely to wander away from the true formant values of the target sentence than at the moderate SNRs. This result might be different if the competing sentence had formants differing greatly from the target sentence.

4. DISCUSSION

The results presented in Figures 2–6 indicate both qualitatively and quantitatively the potential accuracy and robustness of the proposed formant tracking method. These results are only for a single sentence from a male speaker, and suitable algorithm parameters were determined only for this sentence; performance with other sentences and with a female or child speaker may greatly differ. We are currently conducting thorough testing of the algorithm with several target sentences and background noise sources at a range of SNRs.

One problem with the proposed algorithm is that accurate estimation of a formant frequency in voiced speech by LPC requires an analysis window of close to two voic-

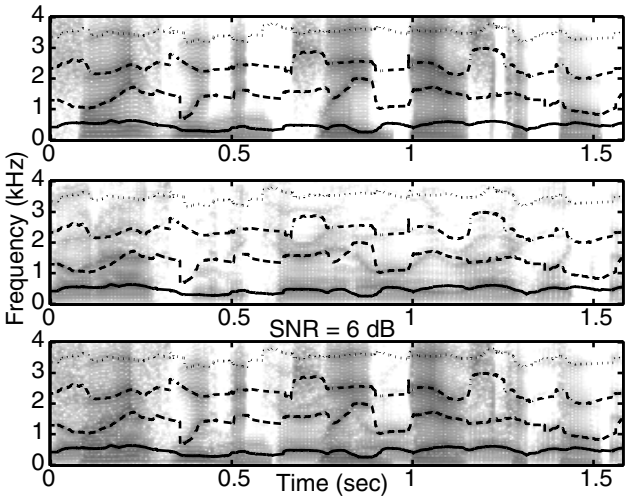

Fig. 5. Formant tracker performance for the synthesized sentence "Five women played basketball" with a competing single speaker ("Don't ask me to carry an oily rag like that," from the TIMIT speech database). Formant tracker estimates are plotted over the spectrograms of the synthesized sentence in the top panel, the competing sentence in the middle panel, and the two sentences combined at an SNR of 6 dB in the bottom panel. The plotting convention is the same as for Fig. 2.

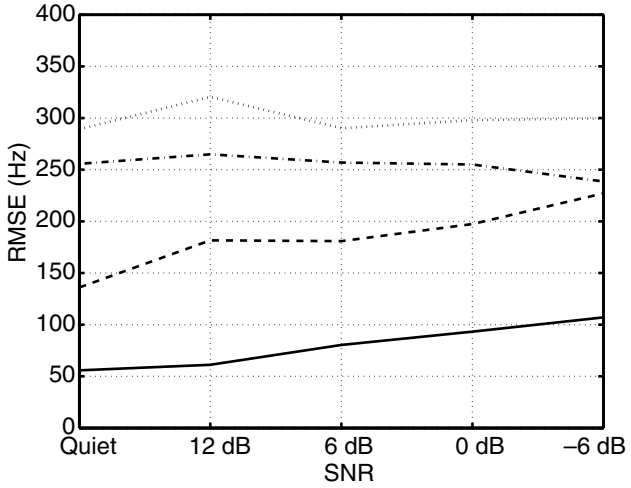

Fig. 6. RMS errors of formant trackers in the presence of a competing speaker as a function of SNR for the synthesized sentence "Five women played basketball." F1 (solid line), F2 (dashed line), F3 (dot-dashed line), F4 (dotted line).

ing pitch periods. A male's voicing pitch is often as low as 100 Hz, requiring a 20 ms analysis window, which results in an average delay of 10 ms. The processing prior to the LPC analysis will add some small additional delay, producing a total delay of 10–15 ms. While this delay may be tolerable for some applications, low-delay alternatives to LPC analysis may be preferable [7]. Difficulty may also arise if background noise or a sudden change in the formant frequencies (e.g., when there is a switch in who is speaking during a conversation) causes the trackers to wander far away from the true formant values. It may therefore be necessary to place limits on the frequency range allowable for each formant tracker.

5. REFERENCES

[1] 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.

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

[3] 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.

[4] 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.

[5] I. C. Bruce, N. V. Karkhanis, E. D. Young, and M. B. Sachs, "Design and physiological testing of dynamic formant enhancement for hearing aids," in *Abstr. 24th ARO Midwinter Meeting*, 2001.

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

[7] J.-H. Chen, "Low-delay coding of speech," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., pp. 209–256. Elsevier Science B.V., Amsterdam, 1995.