MEASURING FINE STRUCTURE IN SPEECH: APPLICATION TO SPEAKER IDENTIFICATION

C. R. Jankowski Jr., T. F. Quatieri, and D. A. Reynolds

MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02173

ABSTRACT

The performance of systems for speaker identification (SID) can be quite good with clean speech, though much lower with degraded speech. Thus it is useful to search for new features for SID, particularly features that are robust over a degraded channel. This paper investigates features that are based on amplitude and frequency modulations of speech formants, high resolution measurement of fundamental frequency and location of "secondary pulses," measured using a high-resolution energy operator. When these features are added to traditional features using an existing SID system with a 168 speaker telephone speech database, SID performance improved by as much as 4% for male speakers and 8.2% for female speakers.

INTRODUCTION[1]

Current systems for speaker identification (SID) can perform well with very clean speech, though performance decreases significantly when speech is recorded under more realistic acoustic conditions, such as over noisy telephone lines [6]. Many of the successful systems to date have relied solely on spectral based features, such as cepstra, which are, however, highly susceptible to degradations from noise and filtering imposed by a communications channel. These parameters are also typically computed over a duration of several pitch periods; therefore they may not adequately measure any fine structure in glottal or vocal tract behavior that would occur more rapidly. In this paper, it is shown that such fine structure can help cue a speaker's identity from degraded speech.

This paper explores the use of three high-resolution non-spectral based features for measuring fine structure in speech: formant AM-FM, "pitch jitter" and "secondary pulses." Formant AM-FM is the modulation of the amplitude and frequency of resonances of the vocal tract. Possible sources of formant AM-FM include rapidly varying parameters of the vocal tract and aerodynamic effects in the vocal tract [7]. Pitch jitter is minute period-to-period variation in fundamental frequency. Lastly, multiple impulses per pitch period, i. e., secondary pulses, are revealed in the envelopes and energies of bandpass-filtered speech, and sometimes the speech waveforms themselves. These secondary pulses may be due to possible multiple glottal excitation sources in the pitch period [2], nonlinear effects in the vocal tract [7], or beating caused by closely spaced formants.

FORMANT AM-FM PARAMETERS

To measure formant AM-FM we use a high-resolution energy operator suggested by Teager and developed by Maragos, Quatieri, and Kaiser [4], [5], [7]. Given an AM-FM sinusoid with a time-varying amplitude and frequency, the Teager energy operator returns a high-resolution energy estimate. Maragos, Quatieri, and Kaiser also developed "energy separation" algorithms that estimate the amplitude and frequency from the Teager energy of a waveform and its derivative. Both the Teager energy operator and the energy separation algorithms require a single (possibly AM-FM) sinusoidal component in the waveform; the output is not meaningful for "multi-component" signals. Thus a signal must be bandpass-filtered around a sinusoidal component before performing energy measurements.

Figure 1 shows a block diagram of a formant AM-FM speech analysis system. Order 19 LPC analysis is first used to find potential locations of speech formants. Using the frequencies and amplitudes from the LPC poles, the first three formants are selected. The speech waveform is then bandpass-filtered around the three formants with Gabor (Gaussian-shaped) filters with bandwidths of 400 Hz. Gabor filters were chosen because of their gradual cutoffs, which reduces incidence of ringing artifacts in the time domain; the frequency response of the filter is also Gaussian-shaped.

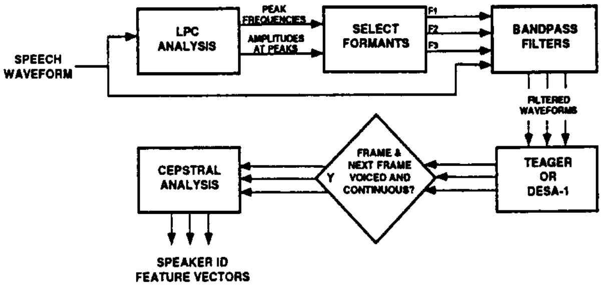

FIGURE 1. Block diagram of speech analysis system for measuring formant AM-FM parameters.

[1] This work was supported by the department of the Air Force. The views expressed are those of the authors and do not reflect the official policy or position of the U. S. Government.

0-7803-2431-5/95 $4.00 © 1995 IEEE

Either the Teager energy operator or the DESA-1 energy separation algorithm [5] is applied to the bandpass waveforms, creating three waveforms of energy, amplitude, or frequency, one waveform for each formant frequency chosen. The sampling rate of these waveforms is the same as the sampling rate of the original waveform.

These analyses are performed in ten millisecond blocks, dividing the speech waveform into frames. For each pair of consecutive 10 ms frames in the utterance, a feature vector is generated if both frames are voiced and the two frames are "similar" (determined by applying a dynamic programming algorithm to the formant frequencies and amplitudes). Because of this similarity constraint, it is not necessary to find highly accurate formant estimates for every speech frame. For each 20 ms segment of the desired quantity (energy, amplitude, or frequency), the d. c. and linear components are removed, the segment is multiplied by a Hamming window, and cepstral coefficients (to be denoted by c[n]) are computed. Mid-quefrency cepstral coefficients are used as feature vectors to the SID system. High-quefrency cepstral coefficients are not used so that pitch information is removed, and low-quefrency cepstra are deleted to remove information about the absolute energy of the formants.

PITCH JITTER AND SECONDARY PULSE PARAMETERS

We also use the Teager energy operator during voiced speech to generate both a high-resolution measure of fundamental frequency[2] and the location of secondary pulses. Figure 2 shows a block diagram of the analysis system. A short-time Fourier transform (STFT) is calculated on the speech signal, and from the STFT magnitude a "low" and a "high" formant are selected. Initially, the low formant is the spectral peak at less than 1500 Hz, and the high formant is the spectral peak above 1500 Hz. These estimates are iteratively refined as in [5].

The speech waveform is then bandpass-filtered around these two formant frequencies using a Gaussian filter, and the Teager energy operator is applied to each filtered waveform. Using the two Teager energy waveforms, four parameters are estimated for each glottal period. The low and high "primary" pulse locations P_l and P_h are the times between pitch periods, as measured using the low and high Teager energies. A pitch measure from a sinusoidal transform system is used as a guide in estimating P_l and P_h; this initial measure is computed over several pitch periods. Likewise, the low and high "secondary" pulse locations S_l and S_h measure the times from the pitch period onset to the secondary pulses. In all cases, these times are measured by finding local maxima in the Teager energy, as indicated graphically in Figure 2.

SPEAKER IDENTIFICATION

This section describes the speaker identification system used for evaluations, and presents the results of several SID experiments.

[2] Using the Teager energy to measure fundamental frequency was also independently performed in [8].

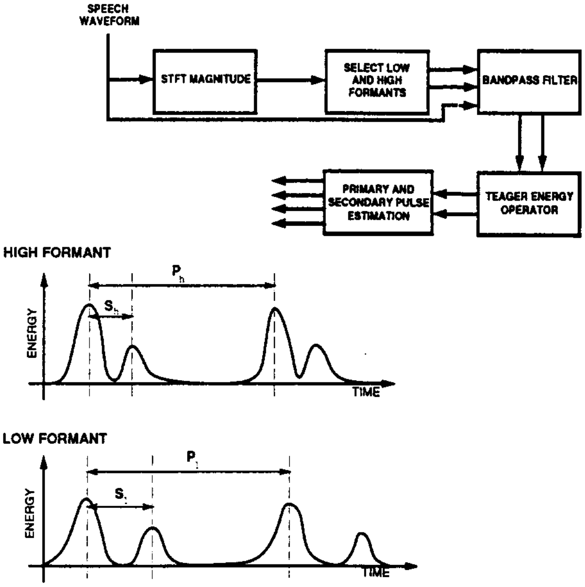

FIGURE 2. Block diagram of speech analysis system for measuring pitch jitter and secondary pulses.

Speaker ID System

A Gaussian mixture model (GMM) SID system [6] is used for all evaluations. During the training phase of this system, all feature vectors from a particular speaker's utterance are used to generate a Gaussian mixture model; i.e., a weighted sum of *M* Normal distributions, each with its own mean, variance, and weight. The number of Gaussian mixtures is set before training.

During testing, feature vectors are assumed independent, so that the probability that a speaker model would generate the string of observations from an utterance is simply the product of the probabilities that the speaker model would generate each of the observations. The flexibility of this system allows the signal processing module to generate feature vectors only during parts of the utterance when the features might be more meaningful and/or robust (e.g., during voiced segments).

Preliminary Experiments with Formant AM-FM

For a first set of SID experiments on clean speech, we used a randomly selected subset of the TIMIT database [1] containing ten male and ten female speakers. To test degraded speech, we also evaluated with the same speakers from the NTIMIT database [3], which is a copy of the TIMIT database that has been transmitted over the telephone network.

For each database and measurement (energy, amplitude, or frequency), a suite of SID experiments was conducted by varying the formant used to compute the cepstral feature vector, the subset of the full cepstral feature vector to pass to the SID system, and the number of Gaussian mixtures used in generating the speaker models. One set of these parameters was chosen for all results: c[9]–c[28] were used from a combination of feature vectors from all three formants, with 16 Gaussian mixtures. Experiments using the frequency estimated from energy separation algorithms produced poor results, so only the Teager energy was used.

The performance obtained was considerably less than can be obtained with standard techniques (e. g., 100% for TIMIT), which is not unexpected since no absolute spectral content is used. For both databases, using cepstra from F1 alone provided best performance of the individual formants (75% and 58% for TIMIT and NTIMIT, respectively). Combining the cepstral vectors from the first three formants resulted in the best overall performance (83% and 75%). While lower than the TIMIT results, the NTIMIT results do not degrade significantly. This is encouraging; one goal of using nonlinear signal processing is to generate features that are robust under noisy and filtered conditions. Based on these and other preliminary experiments, formant AM-FM features and secondary pulse parameters were then combined with mel-cepstral coefficients.

Combining New Features with Mel-Cepstra

To evaluate the effectiveness of a combination of standard and new features, a different database subset of twenty male and twenty female speakers were chosen from NTIMIT. These forty speakers were "difficult" in the sense that they caused the most identification errors when the GMM SID system was run using mel-cepstra alone as the input feature vector. The male subset and female subset were evaluated separately.

Formant AM-FM Results

Mel-cepstra and Teager energy cepstra were combined in the SID system by treating the two feature streams as independent. For each speaker model, the SID system calculated the probability that each of the two feature streams would be generated by the GMM speaker model, and multiplied these probabilities (or added log probabilities) to create the final score. Therefore, mel-cepstra and Teager energy cepstra frames need not be aligned, and it is possible to have a different number of feature vectors in each stream, greatly simplifying front-end processing.

The same parameters are used as in the preliminary experiments, since they consistently produced the best SID performance. With formant AM-FM parameters alone, performance was 27.5% for males and 37.5% for females. Performance of mel-cepstra by itself was 55% for both genders (one binomial standard deviation is 7.9%). The combined performance was 60% for males and 70% for females. It is not surprising that Teager energy cepstra alone does not perform as well as mel-cepstra, considering the lack of absolute spectral information in the Teager energy cepstra. When the Teager energy cepstra are combined with mel-cepstra, however, speaker identification performance improves by 5 percentage points for the male speakers and 15 percentage points for the female speakers. It is interesting that the females improved significantly more; traditionally, female speakers have been more difficult for speech processing systems.

Pitch Jitter and Secondary Pulse Results

We have also combined the four pitch jitter and secondary pulse parameters with mel-cepstra on the same NTIMIT subset. As with the formant AM-FM parameters, the mel-cepstral and Teager-based feature streams were treated as independent streams, and combination required simply multiplying the probabilities from the two streams. This method is especially useful for combining pitch-synchronous features such as pitch jitter and secondary pulse location with frame-based features such as mel-cepstra. 16 Gaussian mixtures were used in the SID system.

Adding mel-cepstra to the high-resolution pitch from the low formant improved performance to 65% and 62% for males and females, respectively. After further adding secondary pulse locations from both formants, performance increased to 68% and 65%. Finally, adding the high-resolution pitch from the high formant improved performance to 70% for males and females.

Our results also showed that using high-resolution fundamental frequency provided 5-7% higher SID performance than using the sinusoidal transform-based fundamental frequency measure computed over several pitch periods, indicating the potential importance of pitch fine structure.

Experiments with Good Performing Speakers

The previous experiments combining formant AM-FM and glottal parameters were conducted with 40 "difficult" NTIMIT speakers. It was important to verify that combining these parameters with mel-cepstra did not degrade performance for speakers with good SID performance. We selected another NTIMIT subset of twenty male and twenty female speakers, for which the mel-cepstral SID system performed perfectly. For both parameter sets and both males and females, performance dropped from 100% to 97.5%; we did not consider this an appreciable loss, and concluded that the new features did not degrade performance on easier speakers.

Combining Features Using a Larger Data Set

The previous experiments combining mel-cepstra with new features were all conducted with 40 testing utterances; both sets of experiments were conducted on either very "bad" or very "good" speakers. In order to increase statistical significance and choose a more representative data set, we chose a larger subset of NTIMIT for further experiments. 168 speakers were selected; 112 males and 56 females.

Formant AM-FM Results

Results for formant AM-FM parameters are shown in Figure 3. One binomial standard deviation for male, female, and total performance is 2.8%, 4.2%, and 2.3%, respectively. For this database, SID was performed on all speakers simultaneously, though there were no gender errors for mel-cepstra or the combination. The same feature and SID system parameters were used as were used in the previous formant AM-FM experiments. Although the new features slightly improved total performance, what is again interesting is that male performance dropped relative to mel-cepstra, but the female score improved by almost two binomial standard deviations. Recall that we also saw greater performance with females with the "difficult" 40 NTIMIT speakers; it should be well worth looking into why the female scores are enhanced by the new features.

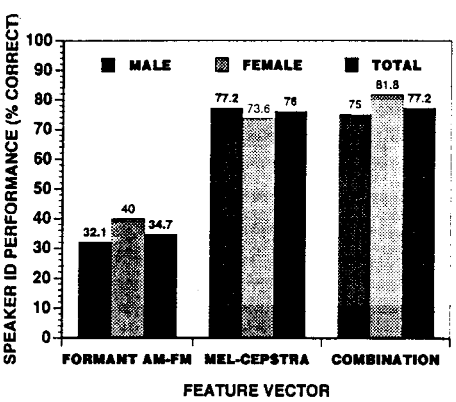

FIGURE 3. Speaker identification performance with formant AM-FM parameters on the NTIMIT 168 speaker subset.

Pitch Jitter and Secondary Pulse Results

Figure 4 shows SID results when combining pitch jitter and secondary pulse parameters with mel-cepstra on the larger data set. 8 Gaussian mixtures were used for all experiments. Compared with mel-cepstra alone, the full combined feature set resulted in a 4% increase with male speakers and no change for the females.

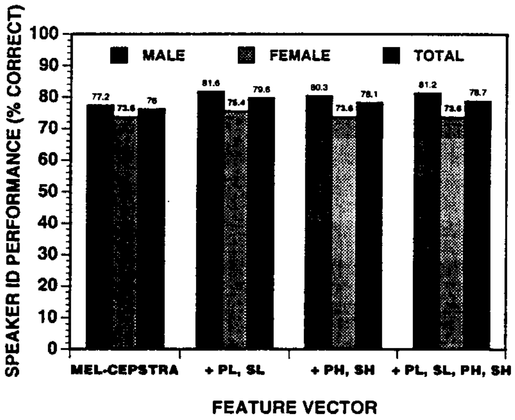

FIGURE 4. Speaker identification performance with pitch jitter and secondary pulse parameters on the NTIMIT 168 speaker subset.

Using the NTIMIT 168 speaker database, mel-cepstral coefficients were also combined with the lower resolution pitch estimate from the sinusoidal transform system, and SID performance was 81.7% for males and 74.5% for females. This performance exceeds that for the full combined pitch jitter/secondary pulse feature vector. This is inconsistent with the results on the 40 difficult NTIMIT speakers, where we saw improved performance with a high-resolution pitch estimate.

CONCLUSIONS

We have described techniques for performing nonlinear high-resolution measurement of speech parameters, and applied these measurements to a speaker identification system. Speaker identification results can be improved when combining these features with spectral-based features, suggesting the potential importance of speech fine structure (previously neglected) for speaker identification. For a 168 speaker NTIMIT data set, formant AM-FM parameters substantially improved SID performance on female speakers, while high-resolution excitation parameters boosted performance for males.

We are continuing to enhance the rather simplistic estimation methods used to generate the new features. We also will investigate the source of the performance difference between male and female speakers. Finally, we will evaluate with other conditions and databases that are more difficult for traditional techniques and might benefit from the introduction of the techniques described in this paper.

ACKNOWLEDGEMENT

The authors acknowledge Petros Maragos, Jim Kaiser, Cliff Weinstein and Jerry O'Leary for useful discussions. The authors also thank the Lincoln Laboratory Innovative Research Program review committee for sponsoring this work.

REFERENCES

[1] W. M. Fisher, G. R. Doddington and K. M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status," *Proc. DARPA Speech Recognition Workshop*, Palo Alto, ARPA ISTO, (1986), pp. 93-99.

[2] J. N. Holmes, "Formant excitation before and after glottal closure," *Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing*, (1976), pp. 39-42.

[3] C. R. Jankowski, A. Kalyanswamy, S. Basson and J. Spitz, "NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database," *Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing*, (1990), pp. 109-112.

[4] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal", *Proc. Intl. Conf. Acous., Speech, and Signal Processing*, Albuquerque, NM, (1990), pp. 381--384.

[5] P. Maragos, J. Kaisér and T. F. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis," *IEEE Trans. Signal Processing* 41, (1993), pp. 3024-3051.

[6] D. A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models," *Proc. ESCA Workshop on Automatic Speaker Recognition*, (1994), pp. 27-30.

[7] H. M. Teager and S. M. Teager, "Evidence for Nonlinear Production Mechanisms in the Vocal Tract," *NATO Advanced Study Institute on Speech Production and Speech Modeling*, Bonas, France, Kluwer, Acad. Pub., (1990), pp. 241-261.

[8] R. K. Whitman and D. M. Etter, "Initial investigation in using an energy operator for pitch estimation," *127th Meeting of the Acoustical Society of America*, Cambridge, Mass, (1994).