FINE STRUCTURE FEATURES FOR SPEAKER IDENTIFICATION

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

MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02173

ABSTRACT[1]

The performance of speaker identification (SID) systems can be improved by the addition of the rapidly varying "fine structure" features of formant amplitude and/or frequency modulation and multiple excitation pulses. This paper shows how the estimation of such fine structure features can be improved further by obtaining better estimates of formant frequency locations and uncovering various sources of error in the feature extraction systems. Most female telephone speech showed "spurious" formants, 'due to distortion in the telephone network. Nevertheless, SID performance was greatest with these spurious formants as formant estimates. A new feature has also been identified which can increase SID performance: cepstral coefficients from noise in the estimated excitation waveform. Finally, statistical tools have been developed to explore the relative importance of features used for SID, with the ultimate goal of uncovering the source of the features that provide SID performance improvement.

BACKGROUND

Speaker identification (SID) algorithms work extremely well with clean speech, but performance degrades considerably with noisy and/or degraded speech such as telephone speech [10]. Our previous work [5] improved SID performance with telephone speech by measuring "fine structure" features, or features that change rapidly, perhaps even within a pitch period. We measure such fine structure using the Teager energy operator [6], a very high-time-resolution estimate of the "energy" (product of amplitude squared and frequency squared) of a single sinusoid.

Two measurement systems have been developed using the Teager energy operator: one is designed to measure vocal tract AM and/or FM of formants, while the other measures glottal excitation characteristics of high-resolution fundamental frequency and the location of "secondary pulses." The formant AM-FM system measures the Teager energy of bandpass-filtered speech waveforms, where the center frequencies of the bandpass filters are the first three formants. Cepstral coefficients of these Teager energies, henceforth referred to as "Teager energy cepstra," are then used as features to an SID system. The glottal system measures the Teager energies of bandpass waveforms from two formants, then uses peak-picking to estimate a high-resolution fundamental frequency and the location of "secondary pulses," or peaks in the Teager energy not occurring around the instant of glottal closure.

A 168 speaker subset of the NTIMIT telephone speech database [4] with 112 male and 56 female speakers was used for all SID experiments. Formant AM-FM and glottal parameters were added to standard mel-cepstra by adding log-likelihoods. With formant AM-FM parameters, SID accuracy degraded from 77.2% to 75% for males, while female accuracy improved from 73.6% to 81.8%. With glottal parameters, male performance rose from 77.2% to 81.2%, while female performance remained the same at 73.6%. Clearly, formant AM-FM parameters help female speakers, while glottal parameters improve performance for males.

FEATURE EXTRACTION

This section describes improvements to both the formant AM-FM and glottal systems by using better formant frequency estimates and investigates sources of error in both systems.

Vocal Tract Parameters

Formant Frequency Estimates

In the original formant AM-FM system, the speech formant frequencies were measured by performing linear predictive (LPC) analysis and choosing pole locations on a frame-by-frame basis. We experimented with the more advanced formant tracker from the ESPS signal processing system [1] to obtain better formant estimates. The ESPS formant tracker first computes a local cost for each mapping from LPC pole locations to formant frequencies, then uses a modified Viterbi algorithm to find the minimum cost mapping between the formants of the previous frame and the formants of the current frame, ensuring formant continuity. Visual inspection showed that the ESPS system was more robust, especially in maintaining continuity of formant tracks. The frame-by-frame LPC analysis had a tendency to flip between labelling a particular track F1 and F2. The original NTIMIT 168 speaker SID experiments were repeated, using the ESPS-generated formants instead of the simple LPC analysis. Male performance improved

1. This work was supported by the Lincoln Laboratory Innovative Research Program, principally sponsored 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.

0-7803-3192-3/96 $5.00©1996 IEEE

by 3.2% to 78.2% with the better formant estimates, which is slightly better than mel-cepstra, while female performance dropped by 0.9% to 80.1%, still significantly better than mel-cepstra. The male performance with ESPS formants is statistically significant (exceeding one binomial standard deviation) relative to results with simple formant estimates.

Sources of Error

At least two possible sources of error have been found in the formant AM-FM system: spurious formants in the NTIMIT database and an error term in the Teager energy cepstra.

Spurious Formants: Spurious "formants" were discovered while using the ESPS formant tracker on voiced[2] frames by comparing the first three formant frequencies of the TIMIT and NTIMIT databases using a dynamic programming matcher, which finds the minimum cost mapping between TIMIT and NTIMIT formants. The cost of a formant mapping is simply the square of the difference between the two frequencies, and a fixed cost is incurred for an NTIMIT formant having no mapping in TIMIT. Table 1 shows the percentage of frames that were matched (all three formants matched), as well as the percentage of frames that showed a spurious F1, F2, and F3. A spurious F1, for instance, would occur if the

	Males	Females
Match	70.0	20.3
Spurious F1	0.0	0.3
Spurious F2	9.8	57.2
Spurious F3	11.5	15.5
Other	8.7	6.7

TABLE 1. Percentage of frames with a match between TIMIT and NTIMIT formants, spurious NTIMIT formants, and other.

"F1" from the formant tracker on NTIMIT did not match any formants in TIMIT, while the NTIMIT "F2" matched the TIMIT "F1." "Other" indicates neither a match nor simply one spurious formant.

Table 1 clearly shows that female speakers have many more frames with spurious formants than do males; we are not sure why this is so. Many spurious formants do not appear to be an artifact of the formant tracker, but actually do exist in the NTIMIT speech signal. Visual inspection suggests that many of the spurious formants are due to harmonic distortion in the NTIMIT waveforms; e. g. the frequency of a spurious "formant" was approximately equal to a low multiple of a "true" formant or a sum or difference of "true" formants. Both of these phenomena are consistent with nonlinear distortion which is known to exist in telephone handsets [4], [11].

These spurious formants imply that in many cases, formants from NTIMIT are not representative of the actual speech formants, and may be artifacts of the speech degradation. In order to discover how valuble better formant estimates might be, SID experiments were conducted using ESPS-generated formants from TIMIT in place of NTIMIT formants as bandpass filter locations in the formant AM-FM system. Curiously, SID accuracy with female speakers dropped by 6.4% to near mel-cepstral performance with TIMIT formants; accuracy with male speakers was not significantly different. For female speakers, i.e. speakers for which the formant AM-FM system showed improvement, measuring "true" formant locations does not result in any SID performance improvement over mel-cepstra alone.

Teager Energy Cepstra Error: A second source of error was discovered in the Teager energy cepstra. The Teager energy operator produces an error term, which is amplified by the cepstral transformation. Due to this error, the Teager energy cepstra are not measureing solely modulations as intended, but also absolute formant frequencies. This component with absolute formant information may be masking modulation information.

Two prominent characteristics are evident in Teager energy cepstra used in the formant AM-FM system: an exponential-like decay, and a modulation pattern on top of the decay. The decay is explained by the bandwidth of the formant; it is the modulation, though, that is the dominant effect in the cepstral window used for the SID experiments (Teager energy cepstral coefficients c[9]–c[28]).

Except in cases of very small (e.g., less than 500 Hz) formant spacing, the *absolute formant location* is the dominant indicator of modulation extent. Consider a bandpass-filtered damped sinusoid;

$$x[n] = (h[n]\cos(\omega n)) \otimes (\alpha^n u[n]\cos(\omega n))$$

where $h[n]\cos(\omega n)$ is the impulse response of the bandpass filter, α^n is the damping function for the sinusoid, and $u[n]$ is the unit step function. According to [8], $x[n]$ is given approximately by:

$$x[n] \cong (h[n] \otimes \alpha^n u[n])\cos(\omega n)$$

i.e., the amplitude of the damped sinusoid is filtered by the baseband version of the bandpass filter. Applying the discrete-time Teager energy operator

$$\Psi(x[n]) = x^2[n] - x[n-1]x[n+1]$$

results in:

$$\Psi(x[n]) = (h[n] \otimes \alpha^n u[n])^2 (\sin\omega)^2 + E(\alpha, h, n)(\cos(2\omega n) + \cos(2\omega))$$

where $E(\alpha, h, n)$ is the amplitude of an *error term* that has a component at both D. C. and at twice the frequency of the sinusoid. This error term indeed results in a modulation of the Teager energy cepstra; the modulation frequency and extent both increase with higher carrier frequency. The logarithmic compression in the cepstral transformation significantly reduces dynamic range, and thus amplifies the effect of the Teager error; various attempts to low-pass-filter the Teager energy resulted in virtually no change in the Teager energy cepstra.

2. Where "voiced" means within a vowel in the TIMIT phonetic labels.

Cepstral coefficients thus could be coding absolute formant information through the Teager energy error. To explore the usefulness of this absolute formant information for SID, a new set of formant AM-FM features were created: before calculating the Teager energy, bandpass-filtered speech was first modulated to a constant frequency (1000 Hz). Absolute carrier frequency information is thus removed from the Teager cepstra. Original formant estimates were used, as opposed to ESPS estimates. Accuracy on males improved to near-mel-cepstral performance, while female accuracy dropped 4.5% to 77.3%, which is still better than mel-cepstra alone but not as good as with the unmodulated Teager cepstra. Thus including absolute formant information harms SID performance on males, while it provides some additional information with females. This is especially interesting considering that the male NTIMIT formant estimates are more frequently close to the "true" formants as estimated from TIMIT.

Glottal Parameters

Formant Frequency Estimates

Glottal paramaters are also sensitive to errors in formant estimates. Initial glottal experiments simply picked peaks in the spectrum as estimates of formant center frequencies. An SID experiment was conducted using LPC pole locations instead of the spectral peaks. Male performance on the NTIMIT 168 speaker subset improved from 81.2% to 85.3%, while female accuracy dropped from 73.6% to 70.9%. Performance thus improved on the speakers for which glottal parameters were already performing well. It is expected that improvement might also be seen with the ESPS formant tracker described above.

Sources of Error

Currently, the glottal system estimates fundamental frequency and secondary pulse locations by looking for peaks in the Teager energy. This simple procedure is prone to error in the estimation of peak times. Another possible source of error is that, for high pitch speakers, secondary pulses could be "masked" by the primary pitch pulses. To improve estimation, we are now experimenting with an analysis-by-synthesis approach as was used by Potomianos [9], in which the error between the Teager energy and a filtered pair of pulses is minimized with respect to the location and amplitude of the secondary pulse. In initial experiments with the analysis-by-synthesis glottal estimation method, with a "difficult"[3] 40 speaker subset of NTIMIT, performance with female speakers was 2.5% higher with the new estimation method. Performance on male speakers was slightly lowered.

SYNCHRONIZED UNVOICING

We have also discovered a new feature, "synchronized unvoicing," that improves SID performance. This feature is based on an estimate of the noisy component of the speech waveform. In many (but not all) cases during voiced speech, this noisy signal has a component that modulates in synchrony with the harmonic component; we call this *synchronized unvoicing*. The noise component also sometimes exhibits secondary pulses; it is possible that this may be yet another clue to speaker identity.

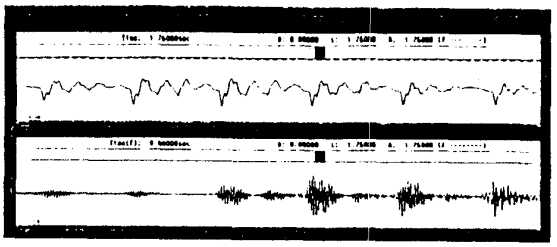

FIGURE 1. Voiced speech segment (top) and corresponding estimated noisy component (4x amplitude expanded).

Using sinusoidal analysis [7], the noisy component of speech was estimated by subtracting an estimate of the harmonic component from the original waveform. The harmonic component was estimated by first measuring pitch; the speech spectral envelope was then sampled at multiples of the estimated fundamental frequency. No harmonics are selected above a voicing-dependent cutoff frequency, so it is possible that the harmonic estimate for unvoiced speech would be zero. The harmonic component of the waveform is created by synthesizing a waveform from only those measured harmonics, while preserving the measured phase of the harmonic components. Figure 1 shows a particular example of the estimated noisy component, where both sychronized noise and secondary pulses are evident. This technique is similar in nature to other work which uses the LPC residual for SID instead of the residual from sinusoidal analysis.

SID experiments were conducted by combining mel-cepstral coefficients from the noisy component with standard mel-cepstra from the entire waveform using the same difficult 40 speaker subset of NTIMIT mentioned above. SID performance improved from 55% to 62.5% for males, and from 55% to 57.5% for females. We plan on replicating these experiments on the entire 168 speaker subset.

RELATIVE IMPORTANCE OF FEATURES

For both the formant AM-FM as well as glottal features, multiple feature streams contributed to SID improvement, e. g., features from various formants. It should be possible to gain insight into the relative importance of features; a tool was therefore developed to automatically generate relative feature weights. The ultimate goal is to discover underlying sources of features; the difference with both systems' performance as a function of speaker gender is of particular interest.

Feature weights were calculated based only on messages that factored in the change in SID score; i.e., those messages that were incorrect with mel-cepstra but correct after adding additional features, or the reverse. We first define a log-likelihood difference $lldif(m_i)$ between the correct speaker and highest scoring competing speaker for message m_i as features were added or removed. $lldif(m_i)$ is positive for a correct message. Since the total score for a message in a SID experiment is computed by add-

3. "Difficult" in the sense that these speakers caused the most SID errors when the 168 speaker subset was evaluated with mel-cepstra alone.

ing the log-likelihoods from the individual feature vectors, the total change in $lldif(m_i)$ after adding all new features is additive. Relative weights of particular formants could thus be computed by normalizing $\Delta lldif_{FN}$ by the total effect of all formants added:

$$w_{FN} = \Delta lldif_{FN} \Big/ \left(\sum_{i=1}^{3} |\Delta lldif_{Fi}| \right)$$

The absolute value in the denominator ensures that weights are normalized to positive and negative effects equally.

We calculated such weights with formant AM-FM parameters. For the males, the magnitude of the weights for messages that improved with formant AM-FM parameters were roughly equal to the weights of messages that degraded with new features. This indicates that each feature vector was equally helpful and harmful, and suggests that removing data from a particular formant should not add any obvious performance gain. F1 was clearly most heavily weighted, followed by F3, and finally F2. For the female messages that were improved, all three formants seem to contribute quite evenly. There was only one female message that was degraded, so there was insufficient data to draw conclusions from these weights.

We also devised a method to calculate the weights of features within feature vectors. In this procedure, a backward search algorithm is used, which removes the feature that causes the least increase (or greatest decrease) in $\Delta lldif$. This feature is removed, the change in $\Delta lldif$ is noted, and the process in continued until only one feature remains. As with the weights calculated above, this procedure is applied only to messages that impacted the change in SID score. Within these messages, the backward search analysis is performed between the correct speaker and all competing speakers that had a higher score than the correct speaker. Feature weights for a given speaker pair are computed by normalizing the change in log likelihood difference after removing a feature, so that the absolute value of the feature weights sum to one. Finally, weights are averaged across speaker pairs and across messages.

This technique was applied to the glottal features, where each feature vector was comprised of three features: secondary pulse location, secondary pulse amplitude, and fundamental frequency. For male speakers, the fundamental frequency was primarily responsible for degrading performance, while the secondary pulse location was most heavily weighted when female performance dropped. For messages where glottal parameters improved performance, fundamental frequency was most important, but secondary pulse location also had a nontrivial weight. Overall, there was little effect of secondary pulse amplitude.

DISCUSSION

In this paper we have shown the importance of good formant estimates for high-performance SID with the new features, and shown some sources of error with both the formant AM-FM and glottal systems. We have also introduced the new feature of synchronized unvoicing. Finally, we have developed a tool for studying the relative importance of features. Using this tool, the ultimate goal is to understand the source of successful features, and perhaps even develop an understanding of the acoustic or physiological bases of these features, so parameters can be measured that result in even better SID performance.

We are also currently beginning to use linear tranformations of features to combine several feature streams, such as mel-cepstra, formant AM-FM, and glottal paramaters. The current method of combination, adding log-likelihoods, assumes independence of feature vectors, which is almost certainly not correct. This independence assumption could be relaxed by simply concatenating feature vectors together, but this results in feature vectors that are too large to train speaker models robustly given the amount of training data. Linear transformations provide a method for optimally reducing the size of these large feature vectors. Already, we have found that linear transformations on mel-cepstral coefficients alone provide significant SID performance improvement.

REFERENCES

[1] Entropic Research Laboratory Inc., "ESPS Programs," (1993).

[2] K. Fukunaga, Introduction to Statistical Pattern Recognition, New York (1972), Academic Press.

[3] J. N. Holmes, "Formant excitation before and after glottal closure," *Proc. Intl. Conf. Acoust., Speech, and Sig. Proc.* (1976) pp. 39-42.

[4] C. R. Jankowski, A. Kalyanswamy, S. Basson, & J. Spitz, "NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database," *Proc. Intl. Conf. Acoust., Speech, and Sig. Proc.*(1990), pp. 109-112.

[5] C. R. Jankowski Jr., T. F. Quatieri, and D. A. Reynolds, "Measuring Fine Structure in Speech: Application to Speaker Identification," *Proc. Intl. Conf. Acoust., Speech, and Sig. Proc.* (1995), pp. 325-328.

[6] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," *Proc. Intl. Conf. Acoust., Speech, and Sig. Proc.* (1990), pp. 381-384.

[7] R. J. McAulay & T. F. Quatieri, "Low Rate Speech Coding Based on the Sinusoidal Speech Model," Chap. 2 in *Advances in Speech Signal Processing*, S. Furui & M. M. Sohndi eds., Marcel Dekker (1991), pp. 165-208.

[8] A. Papoulis, The Fourier Integral and its Applications. New York (1962), McGraw Hill.

[9] A. Potamianos, "Speech Processing Applications Using an AM-FM Modulation Model," Ph. D. thesis, Harvard University (1995).

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

[11] D. Reynolds, M. Zissman, T. Quatieri, G. O'Leary, & B. Carlson, "The Effects of Telephone Transmission Degradations on Speaker Recognition Performance", *Proc. Intl. Conf. Acoust., Speech, and Sig. Proc.* (1995), pp. 329-332.