Signal Modeling Enhancements for Automatic Speech Recognition

Zaki B. Nossair, Peter L. Silsbee, and Stephen A. Zahorian
Dept. of Electrical and Computer Engineering
Old Dominion University, Norfolk, VA 23529

1. Introduction

Obtaining a compact, information-rich representation of the speech signal is an important first step in ASR. A large majority of ASR systems use some form of cepstral coefficients for this purpose. Computation of these cepstral coefficients typically includes several of the following steps: (1) High-frequency preemphasis, using an FIR filter of the form *y(k) = x(k) - ax(k-1)*, with *a* taking values around 0.95; (2) partition of the signal into analysis frames of 20 to 30 ms, spaced 5 to 10 ms apart; (3) computation of ten to forty cepstral coefficients using a cosine transform of the logarithm of the output of a 40-channel triangular filter bank, which is designed to approximate a Bark frequency scale; and (4) Feature vectors are assembled from the instantaneous cepstral values, augmented with some form of dynamic information, e.g. delta-cepstra.

This paper describes several enhancements to this procedure. We show that significant improvements in recognition accuracy can be achieved by modifications in all of these steps, particularly for speech corrupted by noise. In particular, we show that

1. The first order high-frequency pre-emphasis should be replaced by a second order preemphasis of the form:

$$y[n]= 0.3426\, x[n] + 0.4945\, x[n-1] - 0.64\, x[n-1]$$

(coefficients for 16 kHz sampling rate) which gives a broad spectral peak around 3 kHz, and is a better match to the inverse of the equal loudness curve.

2. Better noise immunity is obtained using a greater number of relatively short analysis windows (8-10 ms) and shorter frame spacing (on the order of 2 ms), at least if coupled with the feature representation method described in step 3.

3. 10-15 cepstral coefficients are sufficient to maximize recognition accuracy. These coefficients can be computed as a dot product of the filter-bank or DFT output with precomputed modified cosine basis vectors. The cosine basis vectors may be specified to precisely represent any arbitrary desired warping function.

4. Morphological filtering of the cepstral coefficients increases accuracy and robustness in the presence of noise.

5. Rather than using "instantaneous" cesptra, augmented with delta cepstra, cepstral feature vectors associated with a speech segment (such as an entire phoneme) can be represented using a low-order basis vector representation over time of each vector.

In this paper we show the improvements due to each of these steps individually, and then in combination, using vowel classification experiments with the TIMIT data base. The best result obtained, over 70% accuracy on test vowels, is higher than has previously been reported for the TIMIT data (using the same configuration of training and test data), and significantly higher than that obtained using a "standard" cepstral analysis.

2. Description of the enhancements

A. Preemphasis filter

The usual form for the preemphasis filter is a high-pass FIR filter with a single zero near the origin.

This tends to whiten the speech spectrum as well as emphasizing those frequencies to which the human auditory system is most sensitive. However, this is really only appropriate up to 3 to 4 kHz. Above that range, the sensitivity of human hearing falls off, and there is relatively little linguistic information. Therefore, it is more appropriate to use a second order preemphasis filter. This causes the frequency response to roll off at higher frequencies. This becomes very important in the presence of noise.

B. Analysis windows

ASR system designers have always had to settle for a compromise in their choice of analysis window. To obtain good frequency resolution, a long window is desirable. However, the linguistic importance of some brief transients makes a short window desirable. The usual compromise is to settle for frame lengths of about 20 or 30 ms, with a frame spacing of 5 to 10 ms.

A shorter window, however, is generally sufficient to capture the salient spectral features, *provided the frame spacing is also sufficiently short.* We use an 8 ms window, with 2 ms frame spacing. When the feature trajectories are represented as described in the following subsections, the frequency resolution appears to be very similar to that obtained with the longer window. This effect is shown in Fig. 1, where spectral representations of the same vowel /aa/ are shown, with different window lengths.

C. Cepstral coefficient computation

We compute cepstral coefficients as a dot product of modified basis vectors with the log magnitude spectrum. The method does not require the use of a triangular filter bank--rather, any desired warping function is incorporated in the basis vectors. This allows unlimited flexibility in terms of degree of warping and frequency range selection, using an exact, numerically stable procedure. More details will be presented in the full paper.

D. Morphological filtering

Spectral peaks carry far more information than do spectral valleys. This is especially true in the presence of noise; in this case, the local minima in the spectrum may be completely buried in noise. However, unaltered cepstral coefficients place equal emphasis on spectral valley information and spectral peak information.

We propose a *morphological filtering* of the time-frequency speech signal representation. This class of methods has been used for formant location (Hanson et al. 1994), as well as in speech coding (Hansen, 1991), but has not to our knowledge been applied to the ASR problem. The morphological *dilation* operation can be used to eliminate spectral valleys of a desired width. We use a flat structuring element and smooth over frequency. Let B be a sliding window centered at a point x, and let $c(x)$ be the value of the frequency representation at that point; then the closing operation is defined as:

$$\mathrm{dilate}(x,B) = \max\{y: y = c(x+z), z \quad B\}$$

E. Basis vector expansion

It is of interest to capture not only information about the static spectral features, but also about their trajectories. This is usually accomplished by the use of delta- and delta-delta- parameters. A more robust and comprehensive representation of these trajectories is obtained by expanding each feature over time in a cosine transform. Thus, the set of values for a given feature, over a given time interval, is represented as a set of cosine transform coefficients. Further, these basis vectors are modified such that more emphasis is given to the center region of the segment and less emphasis is given to the end regions.

3. Experiments

We present vowel classification results from experiments incorporating all of the above signal processing enhancements. In the final version of this paper, each enhancement will be evaluated independently, and demonstrated to provide improved accuracy.

A classification experiment, for 16 vowels extracted from TIMIT data base, was conducted using features computed from 151 frames taken from each vowel. These frames span 300 ms centered at the mid point of each vowel. Each frame is of length 8 ms and the frame spacing is 2 ms. Twelve cepstral coefficients were computed for each frame. Five-term cosine expansion was then used for each feature trajectory. Total of 60 coefficients were then used to represent each vowel token. These 60 features were then used as input for a binary pair partition (BPP) neural network system (Rudasi and Zahorian, 1991 and 1992). The results of this experiment were 83.2% for the training data and 70.3 for the test data.

The experiment above was repeated but with noisy speech for the test data. The SNR for the test speech was 6 dB. The test result of this experiment is 48.5%.

The results reported here for both clean and noisy speech are higher than that reported in literature for the same task. For example, Meng and Zue (Meng and Zue, 1990) reported 61.7% for clean speech and 44.5% for noisy speech with SNR=10 dB.

References

J. H. L. Hansen, "Speech Enhancement Employing Adaptive Boundary Detection and Morphological Based Spectral Constraints," *ICASSP-91,* pp. 901-904, 1991.

H. M. Hanson, P. Maragos, and A. Potamianos, "A System for Finding Speech Formants and Modulations via Energy Separation," *IEEE Trans. on Speech and Audio Processing,* vol. 2, no. 3, July 1994.

H. M. Meng and V. W. Zue, "A comparative Study of Acoustic Representations of Speech for Vowel Classification Using Multi-layer Perceptrons," *ICSLP-90*, pp. 1053-1056, 1990.

L. Rudasi and S. A. Zahorian, "Text-Independent Talker Identification with Neural Networks," *ICASSP-91*, pp. 389-392, 1991.

L. Rudasi and S. A. Zahorian, "Text-Independent Speaker Identification using Binary-pair Partitioned Neural Networks," *IJCNN-92*, pp. IV: 679-684, 1992.