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

ABSTRACT

Experiments in modeling speech signals for phoneme classification are described. Enhancements to standard speech processing methods include basis vector representations of dynamic feature trajectories, morphological smoothing (dilation) of spectral features, and the use of many closely spaced, short analysis windows. Results are reported from experiments using the TIMIT database of up to 71.0% correct classification of 16 presegmented vowels in a noise-free environment, and 54.5% correct classification in a 10 dB signal-to-noise ratio environment.

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) partitioning 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) grouping of the instantaneous cepstral values, generally augmented with delta-cepstra, into feature vectors.

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. 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.

2. Morphological filtering (dilation) of the cepstral coefficients increases accuracy and robustness in the presence of noise.

3. 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.

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 for clean speech, 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.

This paper is organized as follows. Sec. 2 contains a description of the basic signal processing methods which are common to all of the experiments reported here. Sec. 3 describes in detail each of the enhancements. In Sec. 4 we describe our experiments and present results which show improvements in vowel classification for each of the modifications.

2. SYSTEM DESCRIPTION

In this section we describe the basic phoneme recognition system. Note that the experiments reported here have used the TIMIT database, hence a sampling rate of 16 kHz should be assumed.

First, the signal is preemphasized with the filter

$$y[n] = x[n] - 0.95x[n-1] + 0.494y[n-1] - 0.64y[n-2] \quad (1)$$

Compared to a more standard single-zero preemphasis, this form has a very similar response at low frequencies, but peaks around 3.2 kHz and rolls off thereafter. This

† Z. B. Nossair is currently with the Dept. of Electrical Engineering, Helwan University, Cairo, Egypt.

0-7803-2431-5/95 $4.00

is more consistent with human hearing and should reduce the deleterious effects of environmental noise with substantial energy at high frequencies. For systems operating at lower sampling rates, this modification may be insignificant (and, of course, different coefficients would be indicated).

The signal is segmented into frames, and a set of cepstral coefficients is calculated for each frame. The samples are first multiplied by a Kaiser window, and then the log of the squared-amplitude of the DFT (for positive frequencies) is calculated. Finally, cepstral coefficients are computed in a manner which directly incorporates bilinear frequency warping [1]. We now describe this method.

It is desired to represent a continuous function $g(\omega)$ as a weighted sum of basis functions $a_i(\omega)$, in a manner which incorporates a warping of the independent variable

$$\omega' = f(\omega) \ . \tag{2}$$

This can always be accomplished for continuous warping functions by "prewarping" the basis vectors, and computing the i^{th} coefficient as

$$c_i = \int g(\omega) a_i \left(f(\omega)\right) \frac{df(\omega)}{d\omega} d\omega \ . \tag{3}$$

We define a new vector $b_i(\omega)$, given by

$$b_i(\omega) = a_i(f(\omega)) \frac{df(\omega)}{d\omega} \ , \tag{4}$$

and compute the integral with respect to the differential element $d\omega'$. If $\omega' = \omega$ at the limits of integration, then the limits are unchanged by the transformation. In discrete frequency, the numerical computation can be performed as a straightforward dot product

$$c_i = \sum_n g(n) b_i(n) \tag{5}$$

where $g(n)$ and $b_i(n)$ are discretized versions of $g(\omega)$ and $b_i(\omega)$, respectively. The desirable result is that the calculation involves a uniform sampling of the function $g(\omega)$. One can consider the kernel of (3) to consist of the product of the warped basis vector $a_i\left(f(\omega)\right)$ and a *window* function $df(\omega)/d\omega$.

We compute cepstral coefficients using this method to incorporate bilinear frequency warping [2]. Here, $g(\omega)$ is the log spectral magnitude to be represented. For bilinear warping, we obtain the warping and window functions:

$$\omega'(\omega) = \omega + 2\tan^{-1} \frac{\alpha \sin\omega}{1 - \alpha\cos\omega} \ , \ 0 \le \omega \le \pi \tag{6}$$

$$\mathrm{w}_f(\omega) = \frac{1-\alpha^2}{1+\alpha^2 - \alpha cos(\omega)} \ , \tag{7}$$

where ω' is the bilinearly warped frequency. In addition, the basis vectors were set to zero for frequencies below about 75 Hz and above about 6 kHz. In all experiments reported here, a value $\alpha = 0.45$ was used.

Note that any desired warping function can easily be incorporated into precomputed basis vectors using this method. This allows unlimited flexibility in terms of degree of warping and frequency range selection, using an exact, numerically stable procedure.

The resulting coefficient vector is classified using a neural network which has been described elsewhere [3, 4]. Details such as frame length and spacing will be discussed in Sec. 3, since these issues are subjects of the current investigation.

3. DESCRIPTION OF THE ENHANCEMENTS

3.1. 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. However, a more robust and comprehensive representation of these trajectories can be 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. Furthermore, 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.

In fact, the same formulation described above in equations (3)—(7) can be used to obtain a set of time basis vectors. However, in this case it is desired to warp *time* in such a way that detail in the center of the segment is emphasized at the expense of detail at the ends of the segment. Specifically, we use a Kaiser window function, specified here for time values normalized to the interval $t \in [0, 1]$:

$$w_t(t) = \frac{J_0\left[\beta\sqrt{1-(2t-1)^2}\right]}{J_0(\beta)} \tag{8}$$

where $J_0(\cdot)$ is a zeroth order Bessel function of the first kind, and $\beta = 8$ in our experiments. The warping function $t'(t)$ is derived numerically from the above equation.

3.2. Morphological filtering

Spectral peaks carry far more information than 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, standard cepstral calculations place equal emphasis on spectral valley information and spectral peak information.

We propose a morphological filtering [5] of the time-frequency speech signal representation. This class of methods has been used for formant location [6], as well as in speech coding [7], 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. Let B be a sliding window centered at a point ω, and let $x(\omega)$ be the value of the frequency representation at that point; then the dilation operation is defined as:

$$\text{dilate}(x(\omega), B) = \max\{x(\omega + \lambda), \lambda \in B\} \ . \quad (9)$$

We have used a flat, three point structuring element such that the dilation operation consists of replacing each sample in the time-frequency plane with the maximum of that sample and its two nearest neighbors in the frequency dimension.

This has two overall effects, when followed by a projection-based smoothing operation such as the basis vector expansion described above. First, the weighting of local spectral peaks is emphasized relative to all other components. Second, local minima (spectral valleys) are entirely eliminated. The result is a spectral representation with an information content which better matches the information which appears to be relevant for human speech perception.

3.3. 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. When the feature trajectories are represented as described in the preceding subsections, the frequency resolution appears to be very similar to that obtained with the longer window. Visual inspection of time-frequency reconstructions show significantly enhanced time resolution with little or no apparent loss in frequency resolution.

For the experiments reported here, we used a frame length of 10 ms., and a frame spacing of 3 ms. This was compared to a frame length of 30 ms. with frame intervals of 10 ms. Noise immunity appears to be enhanced with this type of processing.

4. EXPERIMENTS

A classification experiment, for 16 vowels extracted from the TIMIT data base, was conducted using features computed as described above. Results are summarized in Table 1. Four processing configurations are compared. For each configuration, classification results are reported for the training set and the test set. Performance is evaluated for the test set with two cases: clean speech and noisy speech (10 dB SNR). In both cases, the classifier was trained on clean speech.

The first configuration used only static cepstral features, augmented with delta-cepstra. In this experiment there were fewer features than in the others (12 cepstral coefficients and 12 delta coefficients for a total of 24); essentially a very narrow time slice was analysed. In this configuration, a 30 ms. analysis window was used, with 10 ms. frame spacing.

The other three configurations used data spanning 300 ms centered at the midpoint of each vowel. The time-warped basis vector expansion (fifth order) described in Sec. 3.1 was applied. This gave a total of 60 features (12 cepstral coefficients, the trajectory of each represented by five time coefficients).

In the second experiment, the setup is identical to the first experiment with the exception of the basis vector expansion. Unsurprisingly, since much more data is incorporated into the analysis, results for all three conditions (training set, clean test set, and noisy test set) are much improved.

The third experiment used a configuration identical to the second, except that dilation of the spectrum was performed prior to calculating the warped inverse DCT. This had little effect on the training set or clean-speech results; however, for noisy speech, performance improved from 48.4% correct to 52.3% correct (an 8% reduction in errors).

Finally, we altered the analysis frames to use 10 ms. frames, spaced 3 ms. apart. For this case, a small but significant improvement was seen for both clean and noisy speech, relative to the previous experiment.

5. CONCLUSION

Several enhancements to standard speech signal modeling methods have been presented. Experimental results confirm that morphological dilation of spectral features, followed by time-warped basis vector representation of feature trajectories, can result in improved phoneme classification results. The improvement due to dilation is especially significant in a noisy environment. Significant improvements are also observed when a shorter analysis frame is used in conjunction with a

Experiment	Dynamic Features	Dilation	Frame Length/ Frame Spacing	Percent Correct		
				Training	Test (clean)	Test (noisy)
1	Delta-cepstra	No	30 ms/ 10 ms	64.3	61.3	38.5
2	Basis vectors	No	30 ms/ 10 ms	79.4	69.0	48.4
3	Basis vectors	Yes	30 ms/ 10 ms	79.2	69.2	52.3
4	Basis vectors	Yes	10 ms/ 3 ms	79.7	71.0	54.5

Table 1: Summary of experimental results.

shorter frame spacing.

The results reported here for both clean and noisy speech compare favorably to those reported in literature for the same task, especially compared to standard cepstral analysis. For example, Meng and Zue [8] reported just 61.7% for clean speech and 45.0% for noisy speech (10 dB SNR) using a mel-cepstral analysis. Their best-performing auditory model obtained 66.1% for clean speech and 54.0% for noisy speech. Our results demonstrate that it is possible to obtain performance in this range without the use of complex auditory processing.

Future research will include deeper investigation of the tradeoffs between time and frequency resolution in spectral models of speech; further experiments with morphological methods for improved robustness in the presence of noise; and the inclusion of these methods in a continuous speech recognition system.

6. REFERENCES

[1] P. L. Silsbee, S. A. Zahorian, and Z. B. Nossair, "A warped time-Frequency expansion for speech signal representation," in *IEEE-SP Intl. Symp. on Time-Frequency and Time-Scale Analysis.* IEEE, Oct. 1994.

[2] K.-F. Lee, *Automatic Speech Recognition: the Development of the SPHINX System*, Kluwer Academic Publishers, Boston, 1989.

[3] L. Rudasi and S. Zahorian, "Text-independent talker identification with neural networks," in *Proc. Intl. Conf. Acoust., Speech, Signal Processing*, pp. 389–392, 1991.

[4] L. Rudasi and S. Zahorian, "Text-independent speaker identification using binary-pair partitioned neural networks," in *Intl. Joint Conf. on Neural Networks*, pp. IV: 679–684, 1992.

[5] J. Serra, *Image Analysis and Mathematical Morphology*, Academic Press, NY, 1982.

[6] H. M. Hanson, P. Maragos, and A. Potamianos, "A system for finding speech formants and modulations via energy separation," *IEEE Trans. Speech and Audio Process.*, vol. 2, no. 3, pp. 436–443, July 1994.

[7] J. H. L. Hansen, "Speech enhancement employing adaptive boundary detection and morphological based spectral constraints," in *Proc. Intl. Conf. Acoust., Speech, Signal Processing*, pp. 901–904, 1991.

[8] H. M. Meng and V. W. Zue, "A comparative study of acoustic representations of speech for vowel classification using multi-layer perceptrons," in *Intl. Conf. on Spoken Lang. Process.*, pp. 1053–1056, 1990.