SIGNAL REPRESENTATION FOR ACOUSTIC SEGMENTATION

J. R. Glass and V. W. Zue

Department of Electrical Engineering and Computer Science, and
Research Laboratory of Electronics
Massachusetts Institute of Technology, USA

ABSTRACT – This paper describes an experiment designed to explore the relative merits of different spectral representations for acoustic segmentation. Conventional spectral representations, such as those produced by wideband discrete Fourier transform and by linear prediction, were compared to those based on auditory modeling. Our analysis of 1,000 sentences from 100 speakers indicates that the representations based on auditory modeling appear to be superior.

INTRODUCTION

The task of phonetic recognition can be stated broadly as the determination of the mapping between the acoustic signal and a set of phonological units (e.g., distinctive feature bundles, phonemes, or syllables) used to represent the lexicon. In order to perform such a mapping, it is often desirable to first transform the *continuous* speech signal into a *discrete* set of acoustic segments. This process of acoustic segmentation often makes use of information derived from a short-time spectral representation of the speech signal. Traditionally, the spectral representations are based on discrete Fourier transform or linear prediction. Recently, various signal representations based on models of the human auditory system have been proposed, and anecdotal evidence suggests that such representations may be superior to conventional ones. The purpose of the study reported in this paper was to investigate the relative merits of various spectral representations for acoustic segmentation.

DESCRIPTION OF THE EXPERIMENT

Signal Representation

Five spectral representations were compared. For the *wideband* representation the spectral vector was obtained by applying a 6.7-ms Hamming window to the speech waveform. For the *smoothed narrowband* representation, a 25.6-ms Hamming window was applied to the speech waveform, which was then smoothed with a 2-ms window in the cepstral domain. For the *linear prediction* representation the spectral vector was obtained from a 19th-order LPC analysis on a 25.6-ms segment of Hamming windowed speech. In all three cases, the magnitude spectrum was obtained from a 128-point discrete Fourier transform.

The remaining two spectral representations were derived from an auditory model that incorporates known properties of the human peripheral auditory system, such as critical-band filtering, half-wave rectification, adaptation, saturation, spontaneous response, and synchrony detection (Seneff, 1984; Seneff, 1986). The model consists of three stages. The first performs critical-band filtering with a bank of 39 filters equally spaced on a Bark frequency scale, spanning a frequency range of 130 to 6,400 Hz. The second stage models the transduction process between the hair cells and the neural synapse. The envelope response of the filter outputs corresponds to the mean-rate response of neural firing. The third stage models the synchrony response of the hair cells to their characteristic frequencies. For the *critical-band* representation in our study, the spectral vector was obtained from outputs of the critical-band filters. For the *hair-cell* representation, the spectral vector was obtained from the envelope of the output of the second stage in Seneff's model.

The spectral representations based on auditory modeling have only 39 channel outputs. In the interest of consistency, the remaining three spectral representations were down-sampled accordingly. Thus for each spectral representation, a 39-dimensional spectral vector was computed once every 5 ms. The array of spectral vectors was the only information used for acoustic segmentation.

Acoustic Segmentation Algorithm

Acoustic segmentation was based on an algorithm developed in conjunction with the detection of nasal consonants from continuous speech (Glass & Zue, 1986). Realizing that certain acoustic changes are more significant than others and that the criteria for boundary detection often change as a function of context, we adopted a strategy of measuring the similarity between a given spectral frame and its immediate neighbors. The algorithm moves on a frame-by-frame basis from left to right, and attempts to associate a given frame with its immediate past or future. Specifically, each frame builds up forward and backward cumulative distance contours $D_F(n,i)$ and $D_B(n,-i)$ respectively, with $D(n,i)$ defined as:

$$D(n,i) = \sum_{j=0}^{i} d(n,j)$$

where $d(n,j)$ denotes the Euclidean distances between the feature vector of the current frame, $\vec{v}(n)$, and that of the $n+j^{th}$ frame, $\vec{v}(n+j)$. Then the decision strategy is:

```
Loop  for i from 1 to $I_{max}$
      until $| D_F(n,i) - D_B(n,-i) | > D_{min}$
      finally
      if $D_F(n,i) - D_B(n,-i) > 0$
         then associate frame $n$ to its past
         else associate frame $n$ to its future
```

Thus I_{max} constrains the observation range. Currently this value is set to 50 ms. The threshold D_{min} is a minimum distance threshold indicating when the difference between the two cumulative distance functions is significant enough to form an association. By terminating the search as soon as the threshold is exceeded, the algorithm self-adapts to capture short regions that are acoustically distinct. In addition, the algorithm assigns an association strength, $A(n)$, to each frame, which measures the maximum difference between D_F and D_B in the range of association.

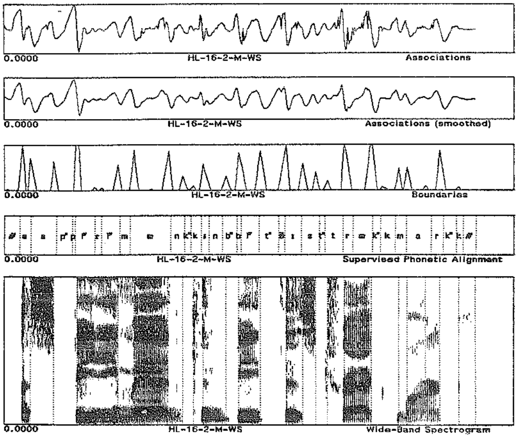

Figure 1. The acoustic segmentation algorithm.

An example of the association waveform is shown in the top part of Figure 1. The positive-to-negative zero-crossings of the waveform correspond to potential acoustic boundaries. To minimize the effect of detecting small and insignificant acoustic changes, this association waveform is smoothed with a Gaussian filter. For the example shown in Figure 1, this smoothed waveform is shown just below the association waveform.

The information in the smoothed association waveform can be captured in the form of a pulse train, also shown for the example in Figure 1. The pulse train provides information not only on the location of the acoustic boundaries, but also on boundary strength (by the height of the pulse) and abruptness (by the width of the pulse). In particular, we found that the height of the pulse is well correlated with the significance of the acoustic change. In other words, smaller pulses typically correspond to insignificant acoustic changes, or false boundaries. Thus it is possible to set a boundary threshold and consider only those spikes whose heights exceed this threshold.

By varying the boundary threshold on the pulse height as well as the amount of smoothing performed on the association waveform, we can control the system's sensitivity to detecting acoustic boundaries. If the sensitivity is set too low, then the system may miss some of the legitimate boundaries. On the other hand, a high sensitivity would tend to insert false boundaries. For the sentence shown in Figure 1, the acoustic boundary locations are superimposed as dotted vertical lines on the spectrogram. By comparing with the time-aligned phonetic transcription above the spectrogram, we see that most of the major acoustic boundaries have been located accurately.

Experimental Conditions

The acoustic segmentation algorithm described above was applied, with varying sensitivities, to each of the spectral representations for a given utterance. The sensitivities were chosen by varying the amount of Gaussian smoothing (using the smallest possible increment for the 5-ms analysis rate) and choosing a boundary threshold that maximized the probability of a boundary being valid. For each experiment, the output of the acoustic segmentation was compared with the time-aligned phonetic transcription of the utterance, and the extra and missed boundaries (*insertions* and *deletions*) were counted. The "best" result for each spectral representation was defined to be the one that minimized the *sum* of the number of inserted and deleted segments.

Databases

We employed two different databases to evaluate the various spectral representations. The first one consisted of a total of 100 phonetically balanced sentences spoken by three male and two female speakers. The second database contained 1,000 phonetically balanced sentences recorded from 50 male and 50 female speakers. All the sentences were previously transcribed phonetically and the transcriptions were time-aligned with acoustic landmarks.

RESULTS AND DISCUSSION

All five of the spectral representations were evaluated using the first database containing over 2,600 phonetic events. The results show that the linear prediction and auditory representations were consistently superior to the discrete Fourier transform representations. Specifically, the critical-band spectral representation was found to be the best (with the lowest total insertion and deletion rate—27%), followed closely by the hair-cell and LPC representations. These representations were consistently better than the discrete Fourier transform representations, by 3 to 4 % on average.

In order to substantiate our preliminary findings, we decided to evaluate the signal representations using the second, larger database of 1,000 sentences containing nearly 29,000 phonetic events. To minimize the amount of computer processing, only the top three representations were evaluated. The results indicate that the auditory representations were consistently better than the linear prediction representation, as shown in Figure 2. The best results were all obtained with a Gaussian smoothing filter of $\sigma = 5$ ms. For all representations, segment insertion is much less likely for this algorithm than segment deletion.

Our results suggest that signal representations based on auditory modeling may be better suited to acoustic segmentation than conventional spectral representations are. While the segmentation errors among various spectral representations did not differ greatly in magnitude, the differences were statistically significant. That the critical-band representation was better than the hair-cell representation in the preliminary evaluation can be attributed to the fact that the first database is rather

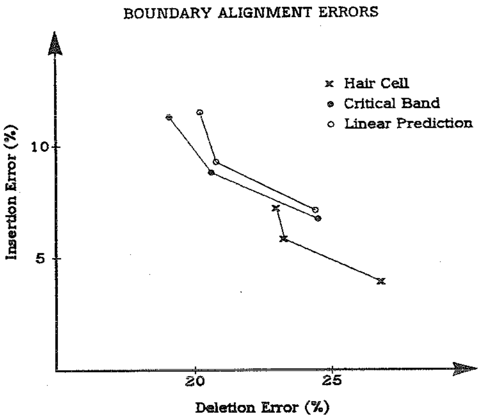

Figure 2. Boundary alignment errors.

small and thus does not provide a good indication of acoustic and interspeaker variabilities. The first database was collected from five speakers using a noise-canceling microphone, whereas the second database was collected from 100 speakers using an omni-directional microphone. Closer examination of the results reveals that the performance of the hair-cell representation actually *improved* for the second database, while the performance degraded for the other two representations.

The hair-cell representation offers a number of advantages. The hair-cell model tends to enhance the onsets and offsets in the critical-band channel outputs. For low-amplitude sounds, the output corresponds to the spontaneous firing of the neurons and is greatly attenuated. These two effects combine to sharpen acoustic boundaries in the speech signal. Figure 3 illustrates the outputs from the three spectral representations. We see that extraneous boundaries are less likely to be observed in the hair-cell representation. We should also note that due to the saturation phenomena, formants in the envelope response appear as broad-band peaks, obscuring detailed differences among similar sounds. As a result, we surmise that this representation may be appropriate for broad phonetic classification as well.

The acoustic segmentation algorithm produces segmentation errors approximately 25% of the time. Closer analysis shows that most of the errors were due to the deletion of subtle acoustic boundaries. Since our approach to phonetic recognition is to utilize the output of the acoustic segmentation algorithm in order to establish robust acoustic landmarks for subsequent detailed analysis, these deleted segments may be recovered at a later stage.

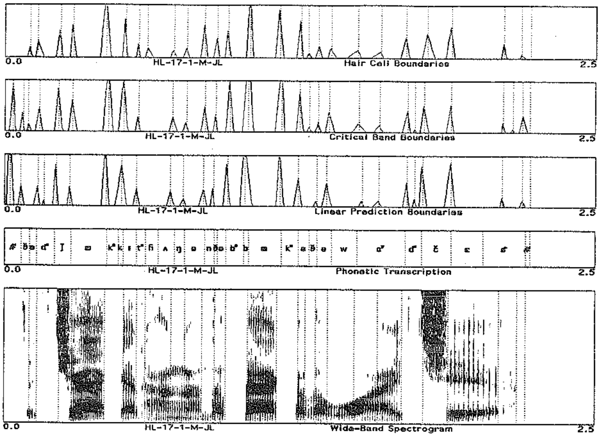

Figure 3. Comparison of different spectral representations.

SUMMARY

In summary, we have investigated the usefulness of five different spectral representations for acoustic segmentation. We found that the spectral representation based on the mean-rate response of an auditory model gives the best performance. This signal representation, together with the acoustic segmentation algorithm, can potentially be used to delineate the speech signal into acoustic regions for further phonetic analysis.

[This research was supported by the Natural Sciences and Engineering Research Council of Canada and by DARPA under Contract N00039-85-C-0254, monitored through Naval Electronic Systems Command.]

REFERENCES

GLASS, J.R., & ZUE, V.W. (1986) "Recognition of Nasal Consonants in American English", Proc. DARPA Speech Recognition Workshop, Report No. SAIC-86/1546, 25–29.

SENEFF, S. (1984) "Pitch and Spectral Estimation of Speech Based on Auditory Synchrony Model," Proc. ICASSP 84, 36.2.1–36.2.4.

SENEFF, S. (1986) "A Computational Model for the Peripheral Auditory System: Application to Speech Recognition Research," Proc. ICASSP 86, 37.8.1–37.8.4.