MULTI-LEVEL ACOUSTIC SEGMENTATION OF CONTINUOUS SPEECH*

James R. Glass and Victor W. Zue

Department of Electrical Engineering and Computer Science, and
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

As part of our goal to better understand the relationship between the speech signal and the underlying phonemic representation, we have developed a procedure that describes the acoustic structure of the signal. Acoustic events are embedded in a multi-level structure in which information ranging from coarse to fine is represented in an organized fashion. An analysis of the acoustic structure, using 500 utterances from 100 different talkers, shows that it captures over 96% of the acoustic-phonetic events of interest with an insertion rate of less than 5%. The paper will describe the signal representation, and the algorithms for determining the acoustic segments and the multi-level structure. Performance results and a comparison with scale-space filtering will also be included. Possible use of this segmental description for automatic speech recognition will be discussed.

INTRODUCTION

The task of phonetic recognition can be stated broadly as the determination of a mapping of the acoustic signal to 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 segments, thus allowing us to focus our attention on important acoustic events. Typically, the *segmentation* process is followed by a *labeling* process, in which the segments are assigned phonetic labels. While this procedure is conceptually straightforward, its implementation has proved to be immensely difficult. Our inability to achieve high-performance phonetic recognition is largely due to the diversity in the acoustic properties of speech sounds. Stop consonants, for example, are produced with abrupt changes in the vocal tract configuration, resulting in distinct acoustic landmarks. Semivowels, on the other hand, are produced with considerably slower articulatory movements, and the associated acoustic transitions are often quite obscure. To complicate matters further, the acoustic properties of phonemes change as a function of context, and the nature of such contextual variation is still poorly understood. As a result, the development of algorithms to locate and classify these phonemes-in-context, or allophones, typically requires intense knowledge engineering.

We are presently exploring a somewhat different approach to phonetic recognition in which the traditional phonetic-level description is bypassed in favor of directly relating the acoustic realizations to the underlying phonemic forms. Our approach is motivated by the observation that a description based on allophones is both incomplete and somewhat arbitrary. Phoneticians traditionally identify a certain number of important allophones for a given phoneme based on their examination of a limited amount of data together with introspective reasoning. With the availability of a large body of data [4], we are now in a position to ascertain whether these categories are acoustically meaningful, and whether additional categories will emerge. Rather than describing the acoustic variations in terms of a set of preconceived units, i.e., allophones, we would like to let the data help us *discover* important regularities. In this line of investigation, the speech signal is transformed into a set of *acoustic* segments, and the relationship between these acoustic segments and the underlying phonemic form is described by a grammar which will be determined through a set of training data.

This paper describes some recent work in acoustic segmentation, as part of the development of a phonetic recognition system. Ideally, the segmentation algorithm should be able to reliably detect abrupt acoustic events such as a stop burst and gradual events such as a vowel to semivowel transition. More importantly, there must exist a coherent framework in which acoustic changes from coarse to fine can be expressed.

SYSTEM DESCRIPTION

The purpose of our acoustic segmentation is to delineate the speech signal into segments that are acoustically homogeneous. Realizing the need to describe varying degrees of acoustic similarity, we have adopted a multi-level representation in which segmentations of different sensitivities are structured in an organized fashion.

*This research was supported by DARPA under Contract N00039-85-C-0254, monitored through Naval Electronic Systems Command.

Determining Acoustic Segments

The algorithm used to establish acoustic segments is a simplified version of the one we developed to detect nasal consonants in continuous speech [1]. This algorithm adopts the strategy of measuring the similarity of each frame to its near neighbors. Similarity is computed by measuring the Euclidean distance between the spectral vector of a given frame and the two frames 10 ms away. Moving on a frame-by-frame basis from left to right, the algorithm associates each frame in the direction, past or future, in which the similarity is greater. Acoustic boundaries are marked whenever the association direction switches from past to future. By varying the parameters of this procedure, we are able to control its sensitivity in detecting acoustic segments in the speech signal. We have chosen to operate with a low deletion rate because mechanisms exist for us to combine segments if necessary at a later stage.

Signal Representation

The algorithms for both acoustic segmentation and classification use the output of an auditory model proposed by Seneff [6]. The model incorporates known properties of the human auditory system, such as critical-band filtering, half-wave rectification, adaptation, saturation, forward masking, spontaneous response, and synchrony detection. The model consists of 40 filters equally spaced on a Bark frequency scale, spanning a frequency range from 130 to 6,400 Hz. For our application, we use the output of the filter channels after they have been processed through a hair-cell/synapse transduction stage. The envelope of the resulting channel outputs corresponds to the "mean rate response" of the auditory nerve fibers. The outputs are represented as a 40-dimensional feature vector, computed once every 5 ms.

We find this representation desirable for several reasons. The transduction stage tends to enhance the onsets and offsets in the critical-band channel outputs. Forward masking will greatly attenuate many low-amplitude sounds because the output falls below the spontaneous firing rate of the nerve fibers. These two effects combine to sharpen acoustic boundaries in the speech signal. Furthermore, due to the saturation phenomena, formants in the envelope response appear as broad-band peaks, obscuring detailed differences among similar sounds, an effect we believe to be advantageous for grouping similar sounds. In a series of experiments comparing various signal representations for acoustic segmentation, we found that, over a wide range of segmentation sensitivities, the auditory-based representation consistently produced the least number of insertion and deletion errors [2].

Multi-Level Description

Our past experience with acoustic segmentation led us to the conclusion that there exists no single level of segmental representation that can adequately describe all the acoustic events of interest. As a result, we have adopted a multi-level representation. We find this representation attractive because it is able to capture both coarse and fine information in one uniform structure. Acoustic-phonetic analysis can then be formulated as a path-finding problem in a highly constrained search space.

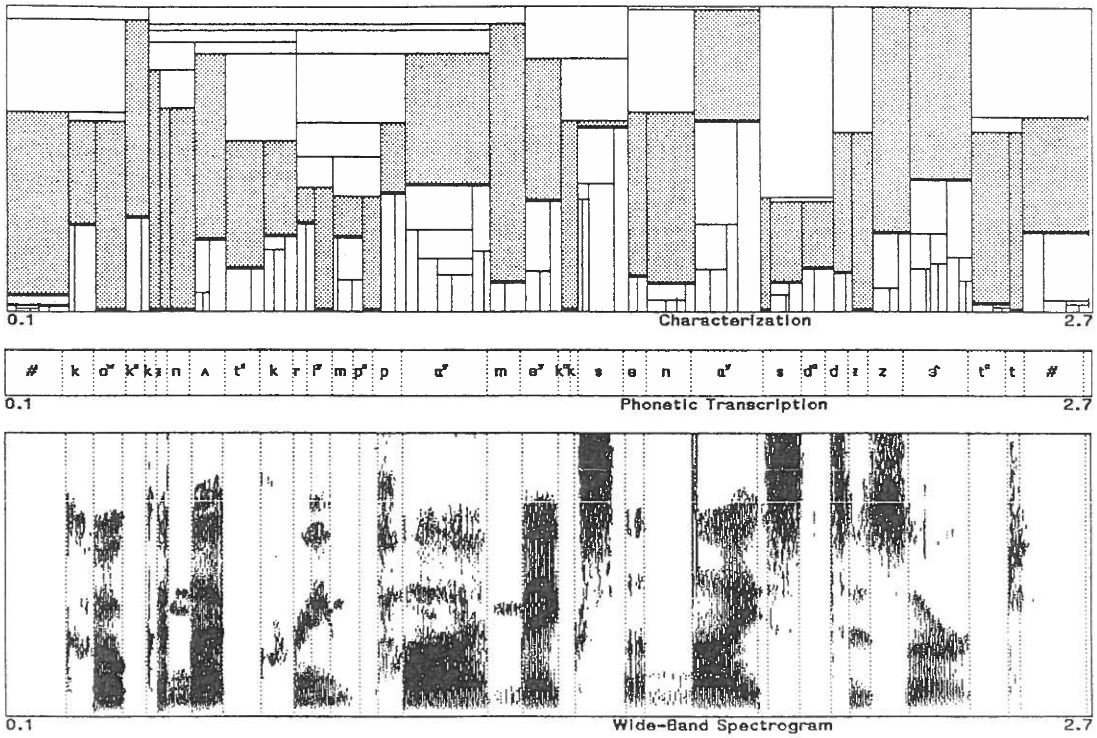
The procedure for obtaining a multi-level representation is similar to that used for finding acoustic segments. First, the algorithm uses all of the proposed segments as "seed regions." Next, each region is associated with either its left or right neighbor using a similarity measure which, in our implementation, is a weighted Euclidean distance measure applied to the average spectral vectors of each region. When two adjacent regions associate with each other, they are merged together to form a single region. This new region subsequently associates itself with one of its neighbors, and the process repeats until the entire utterance is described by a single acoustic event. By keeping track of the distance at which two regions merge into one, the multi-level description can be displayed in a tree-like fashion called a dendrogram, as is illustrated in Figure 1 for the utterance "Coconut cream pie makes a nice dessert." From the bottom towards the top of the dendrogram the acoustic description varies from fine to coarse. The release of the initial /k/, for example, may be considered to be a single acoustic event or a combination of two events (release plus aspiration) depending on the level of detail desired. Similarly, the diphthong /ɑ^{y}/ in the word "nice" may be described as either one acoustic event or two separate ones.

PERFORMANCE EVALUATION

We have evaluated the effectiveness of our multi-level acoustic representation in several ways. First, we developed an algorithm to automatically find the path through the dendrogram which best matched a time-aligned phonetic transcription. An example of such a path is highlighted on the dendrogram in Figure 1. The boundaries along this path are also marked by vertical lines in the spectrogram. We then tabulated the insertion and deletion errors of these paths. Not only should we expect a small number of insertion and deletion errors, the errors should also be acoustically reasonable. Next, we compared the time difference between the boundaries found and the actual boundaries as provided by the transcriptions. Finally, we examined whether correct and incorrect boundaries behave in any reasonable way.

The evaluation was carried out using 500 sentences from the TIMIT database [4]; five sentences each from 100 talkers (69 male and 31 female). These sentences contained nearly 18,500 phones. The best-path alignment procedure gave under 3.5% and 5% deletion and insertion errors, respectively. Closer examination of the errors reveals that the deletions mostly involve acoustic transitions that are not always distinct, such as those between closures and weak stop releases, between vowels and semivowels, between nasals and voiced closures, and between stops and fricatives. In Figure 1, we can see that the boundary between the stop and the frica-

Figure 1: Multi-level Acoustic Segmentation.



tive was deleted in the word "makes". For insertions, it appears that approximately half of the errors occur within the boundaries of a vowel. In Figure 1 there was an insertion between the vowel and the fricative in the word "nice".

Analysis of the time difference between the boundaries found and those provided by the transcription shows that that more than 70% of the boundaries were within 10 ms of each other, and more than 90% were within 20 ms.

Finally, we compared the boundary heights in the dendrogram (as measured by the distance at which the region is merged with one of its neighbors) of valid boundaries to those of invalid boundaries. This comparison is shown in Figure 2. The valid boundaries are typically higher, suggesting that they are more resilient against merging.

DISCUSSION

Our multi-level segmentation procedure is reminiscent of the scale-space filtering idea first proposed by Witkin [8], and investigated by us and others [5,7]. The dendrogram structure, in fact, looks very similar to the interval-tree produced by scale-space filtering. However, there are very important differences between the two procedures. Scale-space filtering produces a multi-level description by uniformly increasing the scale through lowpass filtering, without regard to local context. As a result, at low scales it tends to eliminate short but distinct acoustic events such as stop releases and flaps. In contrast, our procedure merges regions using a local similarity measure. Regions that are acoustically distinct are typically preserved higher in the dendrogram, regardless of their duration. This is illustrated in Figure 3, in which the results of the two multi-level segmentation schemes are compared for the word "coconut." We see, for example, that the short schwa in the second syllable is eliminated at low scales in the scale-space representation, whereas the same vowel and the preceding stop are not combined until much higher in the dendrogram. As an added benefit, our procedure is computationally more efficient, since we represent each region by a single average spectral vector.

The segmentation algorithm uses relational information

Figure 2: Histogram of Boundary Height.

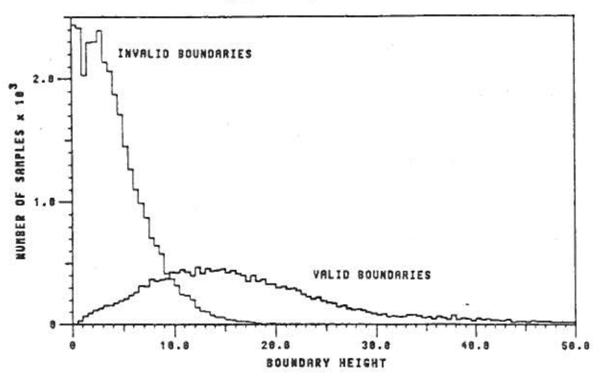
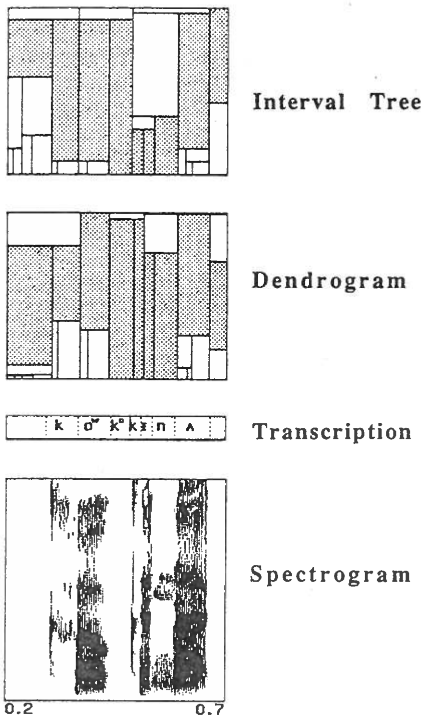


Figure 3: Comparison with Scale-Space Filtering.



within a local context. As a result, we believe that it is fairly insensitive to extra-linguistic factors such as recording conditions, spectral tilt, long term amplitude changes, and background noise. Because these procedures require no training of any kind they are also totally speaker-independent.

Our results on acoustic segmentation suggest that a multi-level representation is potentially very useful. The combined segment insertion and deletion rate of 8.5% is much better than the best result we were able to obtain previously (25%) with a single-level representation, using essentially the same segmentation algorithm and signal representation [2]. Analysis of the errors indicates that most of the deletions occur when the acoustic change is subtle. When a boundary is inserted, it is often the case that significant acoustic change exists, such as within a diphthong or between the frication and aspiration phases of stop releases. Since our objective is to provide an accurate acoustic description of the signal, many of these insertions and deletions perhaps should not be counted as errors.

The dendrogram produces valid boundaries as well as invalid ones, and the distributions of the heights for these two kinds of boundaries are well separated, as shown in Figure 2. The separation becomes even more pronounced when the distributions are conditioned on the general context of the boundary. This type of information lends itself naturally to a probabilistic framework for finding the best path through the dendrogram.

SUMMARY

In summary, we have reported some initial work with acoustic segmentation which we believe can provide a foundation for an eventual phonetic recognition system. By representing the speech signal with a multi-level acoustic description, we are able to capture, and to organize in a meaningful fashion, the majority of acoustic-phonetic events of interest.

The development of the multi-level segmentation procedure is the first step in our development of a phonetic recognition system. We plan to continue our investigation in several directions. First, each region in the dendrogram must be classified into acoustic categories. We have experimented with a hierarchical clustering procedure that produced a small number of robust, acoustic classes that are phonetically meaningful [3]. Next, path-finding algorithms can be used, combining the dendrogram segmentation with the acoustic labels, to select the most likely acoustic interpretation of the utterance. Finally, the dendrogram can be used to help us discover the acoustic regularities of phonemes, taking contextual information into consideration.

REFERENCES

[1] Glass, J.R. and Zue, V.W., "Detection and Recognition of Nasal Consonants in American English," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing,* April, 1986.

[2] Glass, J.R. and Zue, V.W., "Signal Representation for Acoustic Segmentation," *Proc. of the First Australian Conference on Speech Science and Technology,* November, 1986, pp. 124-129.

[3] Glass, J.R. and Zue, V.W., "Acoustic Segmentation and Classification," *Proc. DARPA Speech Recognition Workshop,* Report No. SAIC-87/1644, March, 1987, pp. 38-43.

[4] Lamel, L, Kassel, R., and Seneff, S., "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," *Proc. DARPA Speech Recognition Workshop,* Report No. SAIC-86/1546, February, 1986, pp. 100-109.

[5] Lyon, R.F., "Speech Recognition in Scale Space," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing,* April, 1987.

[6] Seneff, S, "A New Model for the Transduction Stage of the Auditory Periphery," *Proc. DARPA Speech Recognition Workshop,* Report No. SAIC-87/1644, March, 1987, pp. 26-32.

[7] Withgott, M., Bagley, S.C., Lyon, R.F., and Bush, M.A., "Acoustic-Phonetic Segment Classification and Scale-Space Filtering," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing,* April, 1987.

[8] Witkin, A.P.,"Scale Space Filtering: A New Approach to Multi-Scale Description," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing,* April, 1984.