A HMM-based Approach for Segmenting Continuous Speech

B.I.(Raj) Pawate,  
Corporate Research, Development and Engineering,  
Texas Instruments Inc., Dallas, Texas 75265

Eric Dowling,  
Dept. of Electrical Engineering,  
University of Texas at Dallas, Dallas, Texas.

Abstract:  
Speech recognition systems are increasingly utilized in various applications like telephone services where a user places a call by uttering the digits or the name of the person. One of the main problems in this application is the segmentation of the input utterance into speech and non-speech portions. Current approaches typically suffer from two problems. They either incorporate noise as a part of the word to be enrolled or falsely classify a portion of a word as noise. As a result, recognition performance suffers. In this paper, we present another approach to automatically segment continuous speech. To verify our hypothesis, we use a database of 30 speakers whose speech has been recorded over the public switched telephone network. With this database, we benchmark our algorithm against a state of the art approach and show a 4X reduction in the error rate of the recognition system.

Introduction:  
Speech recognition systems are increasingly utilized in various applications today. They are either speaker-dependent or speaker-independent systems; the systems are further classified as isolated where a user utters the words in isolation or continuous where the user utters the words in a continuous normal fashion. For isolated, speaker dependent speech recognizers, the user must first enroll the words. One of the fundamental problems here is how to correctly segment the input speech into speech and non-speech portions; and within the speech portion, how to identify the boundaries of the words.

From these segmented portions, a user-specific template or model is created which is used in subsequent recognition. Recognition performance depends on the accuracy of identifying these speech boundaries within the speech portions of the input utterance. In the case of speaker independent recognition systems, speaker independent templates or models are created off-line by collecting large samples of speech with the target vocabulary words from several speakers. Such large databases are manually segmented to identify the speech boundaries and then transcribed by a trained phonetician. Clearly this task is laborious, error-prone, time consuming and almost impossible if one wishes to create a large vocabulary speaker independent system. Hence the need for a reliable, automatic segmentor which would be used in any speech recognition system.

The segmentation of an input signal is a fundamental problem in speech processing. Typically, an input signal is sampled and several samples are grouped together to form a frame. Sampling rates of 8KHz and frame-sizes varying from 10ms to 30ms are common. These frames...
are processed and represented in the spectral domain by either cepstral coefficients, filter bank amplitudes or principal spectral components. An ideal segmentor would output a sequence of segments s(i) with begin-times b(i) and end-times e(i) and also label the segments. This is shown in figure 1.

Current Approaches:
Over the past years, several algorithms for automatically segmenting an input signal have been proposed. We have classified them into two broad classes - Aposteriori Segmentors and Apriori Segmentors.

Aposteriori Segmentors:
Aposteriori Segmentors do not assume any prior knowledge; they do not make use of models as the other class does. Typically they compute some spectral characteristic of the input signal and attempt to segment. We can further subdivide this class into Endpoint Detectors and Spectral Contour Variation detectors.

Endpoint Detectors:
In this popular approach [1],[2],[3], the RMS frame energy is computed for each frame and an energy profile is created. As shown in figure 2, thresholds are applied to determine when the utterance has begun and when it has ended. There are several advanced variations of this basic theme, in which separate levels of speech and noise-levels are maintained; the speech levels track rising frame energies rapidly and decay slowly when the frame energy starts dropping; the noise level tracks the rising frame energy slowly and tracks the falling frame energy rapidly. Several constraints are applied to flag the end of an utterance.

This scheme allows separation of speech and non-speech portions and is typically used for enrolling words in small-vocabulary, isolated, speaker-dependent speech recognition systems. Within the speech portion, this method does not label the segment i.e. it just identifies the boundaries.

This method has the disadvantage that it can falsely incorporate noise and truncate fricative portions of a speech signal as shown in figure 2. It cannot label the speech portion. It works primarily when the target words are uttered in isolation. This method does not work well in the presence of small signal-to-noise ratios.

Spectral contour variation detectors:
This method [4],[7] assumes that speech consists of more or less stationary portions with transient portions in between. A stationary or steady-state portion is defined as a portion or sequence of frames whose spectra are similar. With this assumption, the segmentation algorithm uses some quantitative criterion of a similarity measure and a searching method for grouping similar consecutive frames. Typical similarity measures include the Itakura-Saito distance for LPC vectors and the Euclidean distance for cepstral vectors or other spectral representations.

Typically, the input signal is processed to eliminate silence at the beginning and end of the file containing the signal. An end-pointer detector previously discussed is used for eliminating silence. A spectral contour is then computed; anchor points are first identified which are defined as those points that correspond to peaks of spectral change. The algorithm then searches for transient portions which are expected to be in the vicinity of these peaks. Once the transient portions have been identified, the steady-state segments are those portions in between the transient segments that exhibit spectral similarity.

Disadvantages of Aposteriori Segmentors:
These methods make a fundamental assumption that they can eliminate silence (or noise, which is assumed to be white). This is not always easy or reliable in low signal-to-noise ratio scenarios.
Second, they do not label the segments. Third, the number of segments found is variable.

**Apriori Segmentors:**

Several segmentors in the literature employ this approach [4],[5],[8],[9],[10]. They assume that the phonetic transcription of the input signal is known and utilize this apriori knowledge. They also utilize models of the segments that they wish to identify. A dynamic programming algorithm, well known in the literature, is then used to align the input signal with the stored reference models. Either dynamic-time-warping algorithms and more recently Viterbi-decoding of a Hidden Markov Model (HMM) is used for time alignment. Most of these segmentors utilize an end-point detector to eliminate noise/silence from the input signal.

These segmentors are classified as Apriori since they assume that a phonetic transcription of the input signal is known prior to segmentation. This transcription is then utilized to invoke the constituent phonemes for time alignment. The other apriori knowledge is the availability of the reference waveforms or models. Note that in contrast to the previous class of segmentors, they identify the boundaries as well as label them.

These segmentors are widely used to segment large databases from which speaker independent models are created.

**Disadvantages of Apriori Segmentors:**

These segmentors are semi-automatic in that the input signal must be transcribed in contrast to the Aposteriori segmentors. Hence they cannot be used to automate segmentation during the enrollment phase of a speaker-dependent recognizer (methods have to been proposed that overcome this limitation by prompting the user to say a target phrase). They assume that a reference phonetic library has been previously created. They also cannot be employed across languages e.g. in going from an English language segmentor to a Dutch language segmentor, the reference models will have to be changed. They focus on explaining the speech portions only and do not fair well in signals with non-speech portions.

Unfortunately, it is difficult to benchmark the individual approaches in the two classes of segmentors and provide some form of a figure of merit since the proponent of each scheme has used his (or her) own specific database for evaluation purposes.

**Proposed Method:**

In this paper we propose another method for segmenting continuous speech. However, before we discuss the algorithm, we will introduce the context in which this approach was developed.

**Context of Algorithm Development:**

A need exists to provide repertory dialing capability to users over the public telephone switched network. Typically, a user enrolls a list of names; after enrollment, the user places a call by just uttering the name e.g. CALL EDITH GODFREY. It is desired to have a system with a minimal training session; ideally, we would like to enroll the name with just one token or sample of speech.

We first used the end-point detector scheme discussed in the Aposteriori class of segmentors. We noted that we could, at best, determine the boundaries of the entire utterance i.e. the segmented portion would contain the name EDITH GODFREY. The algorithm is incapable of segmenting the boundaries between the first and last names - primarily because the utterance is spoken in a continuous fashion. With this approach, we must use a single template or model for the entire name. Subsequent recognition suffers because the user will not necessarily utter the entire name in the same manner that it was enrolled. We usually have to train or update the first pass model with 2 to 3
tokens of speech in order to improve the performance which causes increased training time. Hence the need to seek alternate methods.

Our proposed method does not rely upon frame-energy-profile for segmenting the input speech. Instead, we use a HMM-based speech recognizer for segmentation. Instead of using reference wave forms or models of what we wish to identify or label, we however use explicit models for non-speech signals. For example, we model lip-smack, inhalation, exhalation and background noise. We then recognize the input signal; portions that are not recognized are hypothesized to be the segments corresponding to the word(s) to be enrolled.

Our segmentor schema is shown in figure 3. In this system, we impose certain grammatical constraints as shown in figure 4. The word to be enrolled is called XWORD. F_BG, I_BG and E_BG are non-speech models. F_BG models the kind of signal we see before a user starts to speak. Typically this consists of channel noise, background noise, lip-smack and inhalation. E_BG again models the channel noise, background noise and exhalation. I_BG attempts to model the silence that we notice in between words. The grammar expects to find two words; XWORD is a model that consists of portions of the speech that are not recognized. We model only the durational information for XWORD. For example, we have imposed a constraint that XWORD should have a minimum duration of 100 milliseconds.

In a sense, our proposed method is similar to the segmentors described in the Apriori class of segmentors. However, we differ from this class in four main areas. First, we do not attempt to end-point the input utterance to identify the speech portions. This helps us when the user is attempting to enroll over a noisy channel or when he is in a noisy environment like a busy airport. Second, we do not expect the input utterance to be transcribed; in fact, we have no knowledge of what the input signal contains. (The anchor word "CALL" that we have used in the grammar is optional.) Third, we do not use speech reference waveforms e.g. phonemes as used in the Apriori class of segmentors. Instead, we use non-speech reference wave forms. Hence, our segmentor can be used in training sessions and is independent of the language of the user i.e. the same algorithm and reference models can be used for either English or Dutch input speech. Fourth, instead of focusing on identifying what is recognized, we instead identify portions that are not recognized.

Our approach has the following advantages. Since it does not rely upon frame-energy for end pointing, our segmentor works better with noisy input. Second, since the number of non-speech models (less than ten) are fewer compared to the speech models -phonemes, (typically greater than 10) - the accuracy with which we identify the non-speech portions is better. Third, we show that we can identify inter-word gaps as small as a few frames (tens of milliseconds); this helps us to extract words from continuous speech which improves subsequent speech recognition performance.

Experimental Validation:
In order to validate our approach, we benchmark our algorithm using a database of 30 speakers. A total of 1854 utterances recorded over the public telephone switched network are used. We use a one-pass enrollment scheme i.e. the reference model for the repertory name is created from only one token of speech. Details of the database are shown in table 1. We benchmark our approach against the Aposteriori segmentor (end-point detector). Table 2 shows the results of our experiments. The table shows the performance of the subsequent recognition using the models created from our approach and the end-point detector segmentor. With the frame-energy-based scheme as the baseline, and using only a one-pass enrollment, we show a 4X reduction in the error-rate of the recognition.
Figure 1: Block diagram of proposed segmentation

Figure 2: A fundamental problem

Figure 3: A grammar to control the explanation of the input utterance.
system. Note that the end-pointer detector that we have benchmarked against is a state-of-the-art algorithm fielded today.

Conclusion:
We introduce a novel approach for segmenting continuous speech. We demonstrate a 4X reduction in the error-rate of the recognition system using models created from our segmentor as compared to the error-rate of the recognition system using models created from the frame-energy-based segmentor. Further work needs to be done. For example, we need to compare the accuracy of our automatic segmentor against a manual segmentor - a trained phonetician. We also need to benchmark our segmentor against an Apriori segmentor using phonemes.

REFERENCES:


<table>
<thead>
<tr>
<th>Channel</th>
<th>Public Tel. Network</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Spkrs</td>
<td>30</td>
</tr>
<tr>
<td>Total # utterances</td>
<td>1854</td>
</tr>
</tbody>
</table>

Table 1: Repertory Dialing Database

<table>
<thead>
<tr>
<th>Method</th>
<th>Sentence Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frm.Egy Segmentor</td>
<td>11.1%</td>
</tr>
<tr>
<td>Proposed segmentor</td>
<td>3.1%</td>
</tr>
</tbody>
</table>

Table 2: Benchmarking our proposed segmentor against a state-of-the-art segmentor