BEGIN-END DETECTION USING VOWEL ONSET POINTS

S.R. Mahadeva Prasanna, Jinu Mariam Zachariah and B. Yegnanarayana

Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai-600 036, India
Email: {prasanna,jinu,yegna}@cs.iitm.ernet.in

ABSTRACT
This paper proposes a method for detecting begin and end points of a speech utterance using the knowledge of Vowel Onset Points (VOPs). VOP is defined as the instant at which the onset of vowel takes place. An algorithm for VOP detection in continuous speech is discussed. VOP helps in overcoming the difficulties present in coming up with multiple thresholds followed in most of the existing begin-end detection algorithms. The VOP of the first vowel is used as an anchor point for further analysis to detect the begin of the speech utterance. Similarly, the VOP of the last vowel is used as an anchor point for detecting the end point. The performance of the proposed begin-end detection algorithm is compared with the existing energy-based approach by conducting text-dependent speaker verification experiments. The speaker verification system using the knowledge of VOP for begin-end detection shows a significant improvement in the performance.

1. INTRODUCTION
The need for accurately detecting the begin and end points of a speech utterance is important in many applications like connected digit recognition and text-dependent speaker verification [1, 2]. In all these applications begin-end detection is the first step for selecting the speech regions in the given utterance. Once the begin and end points are located, the feature vectors are extracted from the signal present between these points and used for further processing. During training/enrollment, the extracted feature vectors are stored as reference templates and during testing/verification, the feature vectors are compared with the reference templates by a pattern matching technique like Dynamic Time Warping (DTW) [2, 3]. The decision about the test utterance is made depending on the score obtained from the comparison.

The performance of the system in which the begin-end detection is used as the first stage, depends critically on the accuracy of begin-end detection [2]. The computation process is minimum if the begin and end points are accurately detected. Hence there is a need for an accurate begin-end detection algorithm. There are many algorithms proposed in the literature for the begin-end detection [4, 5]. All of them are mainly based on energy of the speech utterance in one or the other way. In all the algorithms the decision for the begin and the end points of speech is made using multiple thresholds. However coming up with appropriate thresholds is difficult under noisy conditions. Some algorithms also use the knowledge of pitch along with energy for begin-end detection [6]. Detection of pitch in noisy conditions unambiguously is still a difficult task.

Speech is the result of excitation of a time-varying vocal tract system with time-varying excitation. The type of speech sound produced depends on the shape of the vocal tract system as well as type of excitation. The generated speech signal is transmitted through the acoustical media and reaches the listener. The perception system of the listener decodes the speech present in the complex time-frequency patterns with special emphasis to certain regions in the speech signal for
message understanding. There will be significant change in the characteristics of vocal tract system and/or excitation source in these regions and these changes may be considered as events. Thus both from the production and perception point of view, speech may be viewed as sequence of events. Such events may be detected and used them as anchor points for further analysis.

The event-based approach gives an alternative way of analyzing speech, and may also result in better performance as the discriminatory information for analysis is present around the events. Glottal Closure (GC) event and Vowel Onset Point (VOP) event are the production-based events. The GC event is an instant at which closure of vocal folds takes place in a pitch period, and the VOP event is an instant at which onset of vowel takes place [7–11]. In this work knowledge of VOP is used for detecting the begin and end points of speech which helps in overcoming the difficulties present in coming up with multiple thresholds followed in most of the existing begin-end detection algorithms.

The paper is organized as follows: The database used in the present work is described in Section 2. Section 3 discusses the VOP, and an approach to detect VOPs in continuous speech. The performance of VOP detection algorithm is evaluated on the database used in the study. An algorithm for begin-end detection using the knowledge of VOPs is also proposed in this section. The performance of the proposed algorithm is evaluated by conducting text-dependent speaker verification experiments and these studies are discussed in Section 4. The summary of the present work and issues to be addressed further are given in the last section.

2. DESCRIPTION OF DATABASE USED FOR THE STUDY

The objective of this study is to develop a begin-end detection algorithm which is robust against environmental degradations, channel distortions and man made noise during the production of speech, like breathing noise, clicks etc. To study the effect of these conditions, speech data was collected through two different channels one through the microphone and the other through the telephone channel. The quality of the speech signal obtained through these channels differ mainly due to differences in the characteristics of transmission channel. A typical telephone channel has a passband of approximately 300-3300 Hz. Signal energy outside this range are attenuated. In addition to the bandwidth limitation, telephone channel introduces distortion to the spectral characteristics of the speech signal. It also introduces additive noise and glitches.

The data collection was done in the laboratory environment with air conditioning noise and some environmental background noise. Speech data was collected from 30 cooperative speakers for one sentence in Hindi language. The sentence was /vidye sabse baDa dth an hai/, which consists of five words and is approximately 2 secs in duration. Twenty two utterances were collected through microphone and 22 utterances through telephone for each speaker. The speech data collected through the microphone as well as the telephone was sampled at 8000 Hz, and the data was stored as 8 bit wave format.

3. DETECTION OF VOPS IN CONTINUOUS SPEECH

The VOP in each category of sound units may be described in terms of the change in the vocal tract system and the excitation source characteristics [8,11]. For instance in the case of nasal CVs, VOP is described by both the change in the source and system characteristics. The change in the source characteristics is the change in the strength of glottal vibration from initial low value to a high value at the VOP. The change in the system characteristics is from total closure to wide opening. These changes are characterized in the time domain as sudden raise in the strength of the signal at the VOP, and beginning of regular formant structure from the VOP in the frequency domain.

Knowledge of VOPs for different categories of sound units in terms of the changes in the source and system characteristics helps in developing a set of acoustic cues to detect the VOPs, both manually and automatically [11]. For detecting the VOPs, knowledge of the excitation source infor-
mation is derived from the Hilbert envelope of linear prediction (LP) residual [11–13]. It is defined as

$$h(t) = \sqrt{r(t) + r_h(t)}$$  

(1)

where, $r(t)$ is the LP residual of the given signal and $r_h(t)$ is the Hilbert transform of $r(t)$. The Hilbert transform of the signal is obtained by interchanging real and imaginary parts in the Discrete Fourier Transform (DFT) of the signal, and taking its inverse DFT.

A segment of continuous speech, its LP residual and the Hilbert envelope of the LP residual is shown in Figure 1. As shown in the figure, the Hilbert envelope of the LP residual represents the strength of excitation. Also the places of significant change in the strength of excitation are probable candidates for VOPs. From the Hilbert envelope such points may be detected by convolving with a Gabor window having suitable shape [14]. From the study of response of the auditory nerve fibers to tone stimulus, it is found that the course of suppression is longer than the course of excitation [15]. Hence a Gabor window having the shape shown in Figure 2, is suitable for this purpose. This window is generated using the parameter values $\sigma = 100$, where $\sigma$ is spatial spread of the Gabor filter, and $\omega = 0.011$, where $\omega$ is the angular frequency of the sinusoidal component, with a filter length $n = 800$. The result of multiplying the Hilbert envelope of LP residual with the Gabor window, and summing the product for every frame of 100 ms with a shift of one sample for a speech utterance is shown in Figure 3. This is a VOP evidence plot. The peaks are detected using a small positive threshold (5% of maximum evidence, which is chosen to eliminate some peaks occurring due to consonants). The peaks are further validated using some heuristics (see Table 1) to eliminate the spurious ones, and to detect the correct VOPs. The algorithm for VOP detection in continuous speech is given in Table 1. The VOPs detected using the proposed algorithm for a speech utterance collected over a telephone channel is given in Figure 3. The utterance has 8 VOPs, which have been correctly hypothesized and one spurious VOP is also hypothesized. However the spurious VOP is in between the first and last VOPs and will not degrade the performance of Begin-end detection. The heuristics used in the algorithm ensures that no spurious VOPs are hypothesized either at the begin or at the end. This is because the heuristics uses speech knowledge to eliminate spurious ones and the spurious one that may occur either at the beginning or at the end are mainly due to transients like clicks.

![Figure 1: (a) Speech segment, (b) LP residual, and (b) Hilbert envelope of the LP residual.](image)

![Figure 2: Gabor window for $\sigma = 100$, $\omega = 0.011$ and $n = 800$.](image)

In the database considered the utterance has 5 words and 8 VOPs in it. The degradation introduced by the channel, environment and nonspeech things from the speaker like breathing noise, clicks makes it further difficult to unambiguously detect the VOPs. To evaluate the performance, the VOPs for few randomly chosen utterances are manually marked using the knowledge of the strength.
of excitation. The performance of the proposed algorithm in detecting the VOPs is given in Table 2. It can be seen that 95.6% of the VOPs are correctly detected within a deviation of ±30 ms. In the chosen 60 utterances, among the total 480 VOPs, 459 are correctly hypothesized, 21 are missing, and 26 are spurious. Among the missing VOPs, few of them correspond to the first and last VOPs of the utterance. These are the cases when the first vowel and the last vowel strengths are comparable to that of the noise level. This can be attributed to the performance of VOP detection algorithm and can be further improved.

Once the VOPs are detected, the next step is to identify the begin and end points of the speech utterance. To mark the begin and end points of the speech utterance, knowledge of the first and the last VOPs are essential. The first vowel in the utterance may be preceded by a consonant and similarly the last vowel may be followed by a consonant. The consonant is typically of an average length of 200 ms duration. After identifying the first VOP, a point at 200 ms preceding the VOP is marked as the begin point. Using the knowledge of last VOP and VOP Evidence Plot, end of the last vowel is identified. A point at 200 ms from the end of the vowel is marked as the end point. The algorithm for begin-end detection is summarized in Table 3.

4. TEXT-DEPENDENT SPEAKER VERIFICATION STUDIES

To evaluate the effectiveness of the proposed begin-end detection algorithm, text-dependent speaker verification studies were conducted in which begin-end detection plays a crucial role in its performance. The speaker verification experiments were conducted once using the begin-end detection based on the energy and another time using the proposed begin-end detection based on VOPs [2]. The database used for the study is described in Section II.

The speaker verification system consists of the following four stages: preprocessing, feature extraction, training and testing. Preprocessing stage involves begin-end detection, preemphasis and windowing. Speech is converted into a sequence of feature vectors in the feature extraction stage. The feature vectors are used to build reference models during training phase and to verify the claim of the speaker during testing phase.

Endpoint detection is not trivial if the SNR is small. In this study two types of begin-end detection algorithm were used, namely, one based on the energy and the other based on the VOP. The speech is preemphasized to enhance high frequency components. It is further blocked into frames of 20 msec with shift of 6.4 msec and each block is hamming windowed which ensures that the signal discontinuities at the begin and end of each frame is minimized.

After detecting the begin and end points, the speech signal present between these two points is represented as a sequence of feature vectors consisting of spectral components. The shape of the vocal tract contributes to the speaker voice characteristics. That is, shape information of the vocal tract is reflected indirectly in the envelope of the

Figure 3: (a) Speech utterance collected over a telephone channel along with manually marked VOPs, (b) Hilbert envelope of the LP residual, (c) VOP Evidence Plot, and (d) Hypothesized VOPs by the proposed algorithm.
short-term spectra. The spectral information for each frame of the speech signal is represented using the cepstral coefficients and delta-cepstral coefficients. The feature vector consists of 20 weighted cepstral coefficients and 5 delta cepstral coefficients for each frame of the speech signal.

The training phase involves generation of reference templates for each speaker. The templates are cepstral feature vectors for each 20 msec segment of the speech signal. Three reference templates are generated for every speaker and stored for future recall during the testing phase. The reason for generating three reference templates is to take care of the intra-speaker variability of the speaker.

Verification is performed by identification. For this purpose, for each speaker nine other speaker models are used as background speakers. During the testing phase, the test pattern is matched using the DTW technique against the three reference templates of the target speaker, and one of the reference template for each of the nine background speakers, yielding 12 similarity scores. The decision is made based on the ranking of the 12 similarity scores. The 12 similarity scores are stored in ascending order. If the first two ranks are obtained by the claimed speaker templates, then the claim is accepted, else the claim is rejected.

The performance of the system is specified by means of Equal Error Rate (EER), which is the average of the False Rejection (FR) rate and the False Acceptance (FA) rate. Studies were conducted on one sentence for 30 speakers. Twenty two utterances were collected for each speaker. Among these, three utterances were used for creating the reference templates. Out of the remaining 19 utterances, 15 utterances of the speaker were used for conducting the genuine speaker tests. Hence the total number of genuine speaker tests for the 30 speakers is 450 (30 × 15). The remaining four are used for imposter testing during other speaker trails. Imposter tests for a speaker are conducted by giving the utterances of the remaining 29 speakers in the database. For each speaker, two utterances each of the same text of the remaining 29 speakers are taken for testing. Hence, the number of imposter speaker tests is 1740 (30 × 29 × 2).

In order to illustrate the effectiveness of the endpoint detection based on VOP compared to the energy based approach, the system is analyzed using both the methods. Table 4 shows the performance of the text-dependent speaker verification system using short-term spectral features for microphone speech and telephone speech of 30 speakers. False rejection rate indicates the percentage of the number of genuine speakers who have been rejected in the 450 genuine speaker tests conducted, and false acceptance rate indicates percentage of the number of imposter speakers who have been accepted in the 1740 imposter speaker tests.

The system using the energy-based approach for begin-end detection fails more often in the cases where the speech data is noisy, especially in the case of telephone speech. The performance of the system for microphone speech as well as telephone speech data has significantly improved by exploiting the knowledge of the VOPs for begin-end detection.

5. CONCLUSIONS

In this paper an approach for begin-end detection is proposed using the knowledge of VOP. The algorithm is robust for noisy conditions like in the case of telephone channel. Comparison of the proposed begin-end detection algorithm with the existing energy-based approach is made by conducting text-dependent speaker verification experiments. The performance of the speaker verification system is measured in terms of EER which is 2.4 for proposed approach, which is a significant improvement over 5.75 for the energy-based approach.

Presently the VOP detection algorithm uses only excitation source information for hypothesizing VOPs. The algorithm can be improved by incorporating vocal tract system information. Using the knowledge of first and last VOPs, the begin and end points are detected using some heuristics. This can be further improved to come up with a signal processing approach for identifying the begin and end points.
REFERENCES


Table 1: Algorithm for VOP detection in continuous speech

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Downsample input speech to 5 kHz sampling frequency to choose only high signal-to-noise ratio (SNR) regions</td>
</tr>
<tr>
<td>2</td>
<td>Compute LP residual using 20 ms frame size, 10 ms frame shift and LP order as 7</td>
</tr>
<tr>
<td>3</td>
<td>Compute the Hilbert envelope of LP residual and resample it to input sampling frequency</td>
</tr>
<tr>
<td>4</td>
<td>Obtain the VOP Evidence Plot by convolving the Hilbert envelope with the Gabor window for every frame of 100 ms with one sample shift</td>
</tr>
<tr>
<td>5</td>
<td>Choose all peaks above 5% of maximum evidence as probable candidates for VOP</td>
</tr>
<tr>
<td>6</td>
<td>For each peak check for the presence of negative region having at least one value below 5% of the minimum evidence in the VOP Evidence Plot before the next peak. If such region is present at least for duration of 50 ms, then choose such a peak for further validation, else eliminate the peak</td>
</tr>
<tr>
<td>7</td>
<td>In continuous speech two vowels cannot occur in less than 50 ms duration. Hence eliminate peaks which are at a distance less than 50 ms with respect to their next peak</td>
</tr>
<tr>
<td>8</td>
<td>Also in a text-dependent continuous speech case, the VOPs cannot be at a distance more than 500 ms. Hence eliminate the peak if it is at a distance more than 500 ms on either side of neighborhood peaks</td>
</tr>
<tr>
<td>9</td>
<td>Hypothesize the remaining peaks as the VOPs</td>
</tr>
</tbody>
</table>

Table 3: Begin-End detection algorithm

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Detect VOPs in the speech utterance</td>
</tr>
<tr>
<td>2</td>
<td>Mark a point 200 ms prior to the first VOP as the begin point</td>
</tr>
<tr>
<td>3</td>
<td>Detect the end of the last vowel using the last VOP and the VOP evidence plot</td>
</tr>
<tr>
<td>4</td>
<td>Mark a point 200 ms after the vowel region as the end point</td>
</tr>
</tbody>
</table>

Table 4: Performance of the text-dependent speaker verification system with microphone and telephone speech using spectral information. The abbreviations FA, FR, and EER respectively indicate false acceptance rate, false rejection rate and equal error rate.

<table>
<thead>
<tr>
<th>Endpoint detection</th>
<th>Speech Data</th>
<th>FA</th>
<th>FR</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy</td>
<td>Microphone</td>
<td>4.9</td>
<td>2.0</td>
<td>3.45</td>
</tr>
<tr>
<td></td>
<td>Telephone</td>
<td>5.8</td>
<td>5.7</td>
<td>5.75</td>
</tr>
<tr>
<td>VOP</td>
<td>Microphone</td>
<td>3.9</td>
<td>0.2</td>
<td>2.05</td>
</tr>
<tr>
<td></td>
<td>Telephone</td>
<td>4.19</td>
<td>0.6</td>
<td>2.4</td>
</tr>
</tbody>
</table>

Table 2: Performance of VOP detection algorithm evaluated over randomly chosen 60 utterances from the database. The symbols T, H, S, M, and C represent total, hypothesized, spurious, missing, and correctly identified VOPs, respectively.

<table>
<thead>
<tr>
<th>T</th>
<th>H</th>
<th>S</th>
<th>M</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>480</td>
<td>485</td>
<td>26</td>
<td>21</td>
<td>459</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(5.4%)</td>
<td>(4.6%)</td>
<td>(95.6%)</td>
</tr>
</tbody>
</table>