Acoustic Segmentation of Speech Using Zero Time Lifting (ZTL)

RaviShankar Prasad, B. Yegnanarayana

Speech and Vision Labs, IIIT Hyderabad
ravishankar.prasad@research.iiit.ac.in, yegna@iiit.ac.in

Abstract

Automatic segmentation of speech signals has been a constant engineering challenge. Even after the advances with supervised and unsupervised techniques, there still lies a challenge to equal the manually labelled segments. HMM-based segmentation techniques with modifications and corrections have been the state-of-art. These techniques are supervised in nature and thus require availability of large corpus transcribed with phone boundaries. The unsupervised techniques, on the other hand, explore gradients in various spectral and temporal properties of the speech signals. This paper presents a new and unsupervised method based on signal processing techniques to segment the speech signals. A recently developed method known as Zero Time Lifting (ZTL) is used for the analysis of speech signals, which provides fine temporal resolution of the spectral features of the segment being analyzed. It uses the Hilbert envelope of Numerator Group Delay (HNGD) of the signal to highlight its spectral activity. This representation is used to extract high SNR regions of the spectra, which in turn proves to be useful in representation of the production characteristics of the speech signal. Performance of the proposed analysis is at par with the existing baseline systems for unsupervised segmentation.

Index Terms: zero time lifting (ZTL), hibert envelope of numerator group delay (HNGD), zero-frequency filter (ZFF), speech segmentation

1. Introduction

Segmentation of speech signals is a critical pre-processing task for several speech applications. These applications mostly rely on the availability of a corpus containing speech segments usually rich in linguistic and signal contents. Along with this, the corpus is also expected to contain information about its acoustic content with corresponding segment boundaries. The most precise way of maintaining such a corpus with well-defined boundaries of speech and corresponding linguistic units with proper time alignment relies on manual efforts, till date. Manual segmentation and annotation of these databases require a large amount of time and effort, and therefore is a tedious job. Most speech processing applications generally use Hidden Markov Models (HMMs) and utilize the availability of such a corpus for training the models. Automatic speech segmentation techniques find their application with speech recognition, phonetic analysis, speech coding and other related areas of speech technology. A major advantage with automatic speech segmentation techniques is the consistency in their results. Studies indicate that manual labelling and segmentation processes are subjective, and thus may result in significant differences in the transcriptions by different people [1].

Automatic segmentation of speech signals is broadly classified into the explicit or implicit categories [2]. Explicit, being the text-dependent case, where a known phonetic sequence is time aligned against speech segments using a set of phone models or reference patterns. Implicit or text-independent techniques are those where there is no prior knowledge of corresponding phonetic sequence. Therefore, for these techniques, given a continuous speech, there is always non-compliance of the number of phones segments detected, to those actually present. In either of these cases the most commonly used approach includes trained HMM models. This process requires extraction of spectral characteristics of the speech signal followed by the forced alignment of HMM phone or syllable models using Viterbi alignment techniques. These techniques employing phone models to segment are termed as supervised techniques. In earlier attempts, a phone segmentation score of 87% was obtained using HMMs on TIMIT database [3]. Since then, several post-processing techniques have been combined with the HMM-based segmentation techniques to improve the detection scores for boundaries of phonetic segments. A statistical correction method was introduced to attenuate for the errors encountered with the HMM-based segmentation methods [4]. A speaker adaptation technique was also employed for error minimization along with the corrections of hypothesized boundaries to improve the results by almost 10% in both context-dependent and context-independent HMMs. Fusion techniques with multiple features [5] or multiple base segmentation engines [6] using regression methods have also been attempted to improve segmentation results.

The other class of techniques for automatic speech segmentation focuses on identifying changes in the signals in the temporal as well as in the spectral domain. These techniques emphasize on parameterization of speech signal and observing the behavior of these parameters over the entire signal. These are termed as unsupervised techniques and do not require any pre-acquired knowledge with respect to the data. These methods are bottom up approaches, where the lexical context integration is performed after the acoustic processing. In an attempt towards segmentation, the acoustic-phonetic knowledge of various manners and places of articulations was successfully employed with statistical pattern recognition approaches to obtain results comparable to HMM-based methods [7]. A comparative study, on phoneme segment detection using acoustic changes, with manually transcribed boundaries proved that unsupervised techniques are fairly effective and can be improved with infusion of additional information about broader classes based on energy, duration and articulatory cues [8]. An analysis of the hypothesized and missed boundaries with an unsupervised algorithm called the maximum margin clustering (MMC) was performed to improve the results obtained by the algorithm [9]. Another work proposed a delta spectral function (DSF) to represent the gradients in band energy for a specific band to measure the spectral changes [10]. Characterization of the rate of spectral transition to detect phoneme boundaries has also been employed to identify phoneme segment boundaries [11]. All these automatic
represented by the Hilbert envelope of the NGD (HNGD) which characteristics for a signal can be obtained for segments starting at involves the windowing of speech signal using the analysis of speech signals which provides high temporal resolution. This filter is used to multiply the signal with a highly decaying impulse-like window, ensuring high resolution in time. The window function is given by

\[ h[n] = \frac{1}{8 \sin^4(\pi n/N)} \text{, } n = 0, 1, 2, ..., N - 1 \] (2)

This filter is used to multiply the signal \( s[n] \) starting at a reference point \( n=0 \), and this imparts a polynomial-type growth/decay to DFT samples in the frequency domain. The given speech signal is analyzed using ZTL and the frequencies of the dominant peaks from the spectra with their amplitudes are obtained. When these frequencies are plotted along the signal, it was observed that there is clear distinction of various acoustic segments. These frequencies represent the dominant resonances of speech segment being analyzed, and can equivalently be associated with the dimensions of the prominent cavity in the vocal tract responsible for the production of that segment. These resonance peaks are thus called dominant resonant frequencies (DRFs), and are considered as representation of the production characteristics for a speech signal. Figure 2 shows speech signals corresponding to some acoustic events and their DRFs with their respective amplitudes. It can be observed from the figure that DRFs relate to the instantaneous production characteristics for a speech signal and thus provide evidence to identify distinct acoustic segments in the speech signal. The temporal resolution obtained by ZTL helps in plotting DRFs at every sampling instant.

\[ g(\omega) = X_f(\omega)Y_R(\omega) - X_R(\omega)Y_f(\omega) \] (3)

where \( X(\omega) = X_R(\omega) + jX_f(\omega) \) is the DTFT of \( x[n] \) and \( Y(\omega) = Y_R(\omega) + jY_f(\omega) \) is the DTFT of \( x[n] \). The spectrum is represented by the Hilbert envelope of the NGD (HNGD) which has a good resolution around the formants[13]. ZTL analysis involves the windowing of speech signal using \( h[n] \) with a shift of one sample to calculate the spectrum. The spectral characteristics for a signal can be obtained for segments starting at any instant of time, and hence the results can be interpreted as instantaneous spectral features. The energy profile of the ZTL spectra at every instant can be attributed mostly to the signal sample at that instant, and to a few other samples in the vicinity. Figure 1 shows a speech signal for the utterance 'advertising' and the corresponding ZTL spectrum computed using a window length of 10ms at a sampling rate of \( f_s = 16k\text{Hz} \), i.e., for \( N = 160 \), and shifting this window by every sample. The capabilities of the HNGD spectra to provide a high temporal resolution and highlighting the spectral peaks for a given speech signal is well evident from the figure.

The paper is organized as follows: Sections 2 and 3 discusses the development of the ZTL technique and extraction of resonant frequency peaks. Section 4 presents the proposed segmentation method, the database used for evaluation, results and comparison with the existing techniques. Section 5 discusses the possible causes of errors in segment boundary detection as well as the further studies along these lines.

2. Zero Time Liftering : motivation and method

Zero time liftering (ZTL) [12] is a recently proposed method for the analysis of speech signals which provides high temporal resolution maintaining simultaneously a good spectral resolution. ZTL involves multiplying of the speech signal with a highly decaying impulse-like window, ensuring high resolution in time. The window function is given by

\[ \alpha(\text{in}%) = \frac{N_{\text{detected}}}{N_{\text{truth}}} \times 100 \] (1)

where \( N_{\text{detected}} \) refers to the total number of boundaries detected for a given tolerance level and \( N_{\text{truth}} \) is the number of boundaries in the ground truth segmentation. The tolerance level is generally expressed in milliseconds. For a segment boundary to be correct it has to fall within a tolerance window from the boundary marked for the ground truth.

In this paper we propose a method for unsupervised segmentation of speech based on recently developed signal processing techniques. In this method, information about the segments present or any related phone models is not used. The proposed method is independent of language, speaker or context. It has the advantage in terms of representation of vocal tract characteristics of the speech signal. The idea of the zero time liftering (ZTL) [12] method was conceived from a recently developed technique for identifying the glottal closure instants, the zero frequency filtering (ZFF) method. ZTL is an analysis technique which has capabilities to provide good temporal and spectral resolution for speech signals. It highlights the high SNR regions in the spectral domain. These advantages of the ZTL analysis technique served as a motivation for the development of the segmentation algorithm presented in this paper.

The capabilities of the HNGD spectra to provide a high temporal resolution and highlighting the spectral peaks for a given speech signal is well evident from the figure.

3. Obtaining DRF using ZTL and segmentation of speech

ZTL provides an insight to the production characteristics for different acoustic segments of speech signals. Analysis of speech using this method can help in understanding the difference in spectral behavior for segments corresponding to various acoustic events. Figure 1 shows the change in spectra corresponding to various acoustic events. The location of the spectral peaks and their corresponding strengths in the spectra led to the development of representation of speech signal using the most prominent peak. These spectral peaks correspond to high SNR regions which are less affected by environmental factors, and thus are robust in representing the speech signals.

The given speech signal is analyzed using ZTL and the frequency of the dominant peaks from the spectra with their amplitudes are obtained. When these frequencies are plotted along the signal, it was observed that there is clear distinction of various acoustic segments. These frequencies represent the dominant resonances of speech segment being analyzed, and can equivalently be associated with the dimensions of the prominent cavity in the vocal tract responsible for the production of that segment. These resonance peaks are thus called dominant resonant frequencies (DRFs), and are considered as representation of the production characteristics for a speech signal. Figure 2 shows speech signals corresponding to some acoustic events and their DRFs with their respective amplitudes. It can be observed from the figure that DRFs relate to the instantaneous production characteristics for a speech signal and thus provide evidence to identify distinct acoustic segments in the speech signal. The temporal resolution obtained by ZTL helps in plotting DRFs at every sampling instant.
3.1. Distinct acoustic events and respective DRF behavior

The DRFs of the speech signal help demarcate boundaries for distinct acoustic events in the speech signal with good accuracy. The acoustic events correspond to the change in the shape of the vocal tract which produce different phonetic classes of sounds. Multi-dimensional representation of the source and system responses are used to build models to learn the changes during the production of the speech signal. These learning processes are supervised in nature, and thus require a large number of instances for each of such transitions. Employing DRFs on the other hand doesn’t require any training but just the pre-acquired knowledge of production mechanism for different classes of sounds. Voiced segments, for instance, are produced with a cavity which is wide open without creating any constriction in the vocal tract. Whereas obstruents are produced by creating constrictions at various locations using different articulators, and this gives rise to different cavity shapes for these sounds. Such cavities, being smaller in length compared to the voiced sounds results an increase in the frequency of resonance. Nasals are other class of sounds which are produced by the coupling of the vocal and nasal tracts by lowering of the velum. This results in a cavity length longer as compared to vowel sounds and thus results in energy distribution in multiple frequency components with a short onset and offset. These and many more acoustic segment transitions are captured efficiently using DRF representation. Experiments were conducted to test the consistency of DRFs across different utterances and speakers. In the case of clean speech, DRFs proved to be quite robust and consistent in representing the production characteristics for the speech signals.

4. Database, method and segmentation results

The segmentation problem requires ZTL analysis to be performed on the given speech signals. An analysis with different window length, $N = 2.5, 10$ and $20\text{ms}$ as given in Eq (2), was carried out for speech, and DRFs were obtained from the corresponding HNGD spectra. On examination of the DRF patterns, we choose $N = 10\text{ms}$ where the DRF representations appear smooth, and segmentation can be done easily.

The algorithm to perform the segmentation tries to identify the changes in the acoustic properties of the signal. To segment the speech signal, we first differentiate between obstruent and sonorant regions based on the characteristics of of the respective DRFs as explained in section 3.1. Further, a 3-point median filtering is performed in the sonorant regions to smooth out the DRF curves. Changes in vocal tract cavity shape within sonorant regions can be identified with a transition in location of DRFs. The transition parameter $f_{tr}$ controls the number of segment boundaries being generated by the proposed algorithm. It is observed that a range of $f_{tr} = 30$ to $120\text{Hz}$ provides almost similar values of $\beta$. Multiple boundaries occurring within a window of $20\text{ms}$ are then merged to one to avoid ambiguity.

![Figure 2: Different acoustic segment of speech and corresponding DRFs along with their magnitude. Plots ‘a’ show speech segments for voiced, nasal, friction and stop respective and plots ‘b’ and ‘c’ show the corresponding DRFs and respective spectral magnitude.](image)

![Figure 3: Segmentation results based on DRF representation of speech signal. Plot ‘a’ shows a speech signal and the corresponding manually transcribed boundaries. Plot ‘b’ shows the corresponding DRFs and boundaries obtained with this representation. Plot ‘c’ shows the sonorant regions with corresponding boundaries with 3-point median filtered DRF curves.](image)

The segmentation algorithm was evaluated on a subset chosen from the TIMIT database [14]. TIMIT is the most widely used corpus for phone segmentation task. It consists of microphone quality recordings of 630 American-English speakers (10 sentences per speaker), with sampling frequency $16\text{kHz}$ and resolution $16\text{-bit}$. The chosen subset contains 182 sentences from TIMIT dataset, uttered by an equal number of male and female speakers. Figure 3 shows one such waveform with the manual transcription boundaries and the corresponding DRF representation of the signal with the boundaries generated by the proposed method. We can see that DRFs represents the changes in production characteristics in the given signal.

The performance of segmentation is reported by comparing the algorithmic segmentation with manual labels provided with the TIMIT database in terms of CDR($\alpha$). A tolerance window of $20\text{ms}$ is generally chosen to report the performances [8]. We computed the results for the segmentation using DRFs for tolerance levels of $5\text{ms}$, $10\text{ms}$, $15\text{ms}$ and $20\text{ms}$, which are shown in Table 1.

Another parameter comparing the performance of segmentation algorithms is the over-segmentation rate (OSR) $\beta$, for the algorithm. The percentage of over-segmentation is given as

$$\beta(\text{in}\%) = \left( \frac{N_{\text{detected}}}{N_{\text{truth}}} - 1 \right) \times 100,$$

(4)
Table 1: Performance of segmentation using DRFs on TIMIT database for different tolerance levels. The CDR is expressed in % and tolerance is expressed in ms.

<table>
<thead>
<tr>
<th>Tolerance</th>
<th>5ms</th>
<th>10ms</th>
<th>15ms</th>
<th>20ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>33.7</td>
<td>58.4</td>
<td>70.9</td>
<td>79.6</td>
</tr>
</tbody>
</table>

Table 2: Comparison of segmentation results for a tolerance level of 20ms. The CDR(α) is expressed in % and tolerance is expressed in ms.

<table>
<thead>
<tr>
<th>Segmentation (supervised)</th>
<th>α</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM+SVR</td>
<td>88.18</td>
</tr>
<tr>
<td>(HMM+SVM)$_1$</td>
<td>94.9</td>
</tr>
<tr>
<td>(HMM+SVM)$_2$</td>
<td>95.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Segmentation (unsupervised)</th>
<th>α</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMC</td>
<td>67.9</td>
</tr>
<tr>
<td>AP$_{seg}$</td>
<td>77.2</td>
</tr>
<tr>
<td>DSF</td>
<td>77.2</td>
</tr>
<tr>
<td>DRF</td>
<td>79.6</td>
</tr>
</tbody>
</table>

where a negative $\beta$ suggests the under-segmentation rate. Silence regions in speech signals have been reported as problematic for unsupervised speech segmentation algorithms [8] and therefore we ignored the segment boundaries within the silence regions for calculating $\beta$. For the rest of the segments boundaries, the proposed algorithm gives a $\beta$ of around 11%. As stated in the previous sections, the segment boundaries detected using DRFs correspond to acoustic changes, which sometimes may not correspond to any of the manually transcribed phoneme boundaries.

Table 2 shows the results obtained for the proposed method in comparison with results obtained by other methods over the same database. The comparison is made with respect to supervised as well as unsupervised segmentation techniques. For instance, HMM + SVR employs multiple base segmentation engines (BSes) which are implemented with HMMs trained on different parameterization methods such as MFCC, LPC, MFCC, PLP etc. and using support vector regression for boundary fusion. This method is explicit in nature, whereas both HMM + SVM methods are implicit, which is basically a segmentation method used for TTS systems. The (HMM+SVN)$_1$ and (HMM+SVN)$_2$ are similar models trained on the TIMIT database but tested on TIMIT and TTS datasets, respectively. They use trained phone models as HMMs and perform SVM (support vector machine) based refinement of local boundaries. Unsupervised methods like the maximum margin clustering, MMC, is a kernel based unsupervised form of SVMs which maximizes the separation margin between a set of unlabelled feature vectors. STM (Spectral transition measure) measures the magnitude of the spectral rate of change and DSF (delta spectral function) represents variation of band energy for a specific band for each frame. Phoneme transitions are usually reflected as peaks of such functions. The AP$_{seg}$ method includes acoustic-phonetic features such as zero crossing rate ZCR, energy onset and offsets and formant energy ratio along with statistical learning methods to detect segment boundaries. When compared to the unsupervised methods, the proposed DRF based method gives better performance with a low $\beta$. There still lies some gap between performances of supervised and unsupervised segmentation methods which can be overcome by further refinements.

5. Error analysis and conclusions

The boundaries obtained by segmentation using the DRFs helps demarcating the acoustic events for a signal. These events boundaries signify the transition in vocal tract characteristics during the production of speech and ZTL analysis helps in marking these accurately. Yet Table 1 shows a low value of $\alpha$ at 5ms and 10ms tolerance levels. An error analysis was carried out over the segment boundaries obtained with DRFs and observation of the signal characteristics in the vicinity of these boundaries. This analysis suggests that the manual transcriptions might be wrongly placed in some cases. The event boundaries demarcation process with DRFs is based on the extraction of acoustic properties of the signal, and therefore are likely to be more accurate.

There are several advantages with the proposed method for representing the speech signals. This method is unsupervised, as it has no requirements in terms of the acquiring knowledge and learning from examples. The results obtained prove the capability of DRFs as consistent evidence to identify the acoustic boundaries of speech signals. Furthermore, this method is independent of language, gender and several other speaker and corpus based dependencies. The representation provided by DRFs for the segments in speech signal also helps in visualizing the boundaries manually, primarily for annotation purposes. Resolution of automatic methods depend on their frame sizes which is generally 10-20 ms in size, whereas ZTL has a high resolution comparable to manual segmentation process.

Future work is planned to incorporate other speech production based features, which together with DRFs can help in improving the segmentation performance. It is also proposed to automatically label the acoustic segments into different categories.

6. References


