Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 144–152
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Phoneme Boundary Analysis using Multiway Geometric Properties of Waveform Trajectories

Parabattina Bhagath, Pradip K. Das
Department of Computer Science and Engineering
Indian Institute of Technology Guwahati, Assam state, India
bhagath.2014, pkdas @ iitg.ac.in

Abstract

Automatic phoneme segmentation is an important problem in speech processing. It helps in improving the recognition quality by providing a proper segmentation information of phonemes or phonetic units. Inappropriate segmentation may lead to recognition accuracy falloff. The problem is essential not only for recognition but also for annotation purpose. In general, segmentation algorithms rely on large datasets for training where data is observed to find the patterns among them. But this process is not straight forward for languages that are under resourced because of less availability of datasets. In this paper, we propose a method that uses geometrical properties of waveform trajectory where intra signal variations are studied and used for segmentation. The method does not rely on large datasets for training. The geometric properties are extracted as linear structural changes in a raw waveform. The methods and findings of the study are presented.

Keywords:
Phoneme boundary,
Peak attributes,
Valley attributes,
Geometrical properties,
CCA.

1. Introduction

Speech recognition is a well-known area that deals with the understanding of spoken units (words, sentences) that has been spoken. It is fair to say that a speech recognition system should be equipped with a good segmentation procedure. A segmentation algorithm essentially identifies the boundaries between two consecutive phonemes in a word or sentence. For an input signal $S[n]$, a segmentation algorithm provides a set of points $b_0, b_1, ..., b_n$ such that the regions separated by these points belong to different phonemes. Phoneme segmentation has to be looked carefully to improve recognition accuracy. This problem has been studied by researchers in different ways.

A conventional segmentation procedure relies on features that can help to understand the changes in speech signals. This information is further processed by any modeling technique of choice to identify the required boundaries. So it is a common practice that a boundary detection involves some feature extraction methods. In literature, a variety of these techniques have been used for this purpose. They are generally categorized as temporal and spectral. Temporal features (Ali et al., 1999) like energy, ZCR (Zero Crossing Rate), Pitch period, LPCCs (Linear Predictive Cepstral Coefficients) are useful in understanding temporal changes in a speech signal. Spectral features like MFCCs (Mel Frequency Cepstral Coefficients), formants, etc. are used to analyze frequency components in a signal. In addition to these, phonetic studies are proven to be helpful in the segmentation task. Research has shown that HMM based systems alone are not sufficient to understand the temporal changes effectively (Yan et al., 2006). It is understood from the studies that structural processing methods are superior to conventional methods in capturing temporal patterns of the signals (Deng and Strik, 2007). Modeling speech trajectory properties are useful to capture the temporal dynamics over the signal which can help to develop dynamic speech models (Liu and Sim, 2012). Even though these methods are effective in capturing temporal dynamics, computational cost and the need for a vast dataset are not relaxed. The present work aimed to develop a reasonable method for phoneme segmentation by incorporating the structural properties of a waveform which can work well on small-sized datasets. The proposed method uses attributes of waveform trajectories to identify the appropriate boundary points using Canonical Correlation Analysis (CCA).

The paper is organized as follows: The next section describes trajectory methods that were used for pattern analysis. Section 3 gives an overview of the CCA method. Section 4 explains the proposed approach for segmentation. The data and experimental setup is described in Section 5. Section 6 explains the results found in the study and Section 7 concludes the paper.

2. Trajectories for Pattern Analysis

In an Euclidean space, a trajectory is defined as a curve that is formed by the observation of the path that a moving object makes. The points in the path are characterized as ordered positional points. Trajectories that were initially known as Linear Trajectory Segmental Models (LTSMs) have been used to analyze speech signals for past 3 decades (Russell and Holmes, 1997). The need for LTSMs point back to the independence assumption in HMM systems. The basic underlying idea in these systems is understanding and equipping models with the knowledge of temporal patterns across segments of a signal. These dynamic features help in overcoming the problem of independence assumption in HMM systems. In LTSMs, each segment is treated as a homogeneous unit that helps in capturing the inter-segmental dependencies too (Yifan Gong, 1997). Trajectories are suitable in pattern analysis for two reasons (Siohan and Yifan Gong, 1996):

1. A speech trajectory is also influenced by the context
2. Trajectories formed by different phonetic units can create independent clusters based on the contextual information

However the models that are based on HMM are suitable for large vocabulary speech recognition (Mitra et al., 2013). Trajectories are not only used for speech signal analysis, but also for pattern analysis in different areas like road network (Atev et al., 2010), databases (Jeung et al., 2008), traffic management, etc.

In general, a trajectory contains vital information like spatiality and temporal patterns of an object. There can be different ways of treating trajectories as segments sequence and points sequence. The similarities in these entities can contribute to crucial knowledge. The similarity metrics to measure the affinity vary on the kind of trajectory. The effectiveness of the comparison method depends on the underlying components that the trajectory represents. Huanhuan et al. proposed a fusion based similarity method for traffic flow patterns (Li et al., 2018). The method combines different techniques like Merge Distance (MD), Multi Dimensional Scaling (MDS) and Density Based Spatial Clustering of applications with noise (DBSCAN) to identify traffic flow patterns and customary routes from vehicle movements. One of the fusion techniques is given by Equation 1.

$$MMTD(t_1, t_2) = 1 - (w_1, w_2)\begin{pmatrix} dist_1(t_1, t_2) \\ dist_2(t_1, t_2) \end{pmatrix} \quad (1)$$

where $dist_1$ and $dist_2$ are different similarity measurements and each measure is treated with unequal weightages. MMTD is maximum-minimum trajectory distance (Xiao et al., 2019) (Lin et al., 2019). The present work uses CCA as measurement metric which is described in the next section.

3. Canonical Correlation Analysis (CCA)

Canonical correlation analysis (CCA) was introduced by Hoteling for multi-variate analysis. It helps to find the relation between multiple variables simultaneously that makes analysis easy. The fundamental step in CCA is to find a set of transforming variables that can transform variables such that the transformation in the corresponding new coordinates is maximally correlated. In the process, a set of variables called as canonical weights are used. The solution to this is computationally expensive and time consuming. Therefore, it is convenient to solve the problem as an eigen value problem. The objective function to solve CCA for two variables x and y can be expressed by Equation 2:

$$C = \begin{pmatrix} 0 & C_{xy} \\ C_{yx} & 0 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \rho^2 \begin{pmatrix} C_{xx} & 0 \\ 0 & C_{xx} \end{pmatrix} \quad (2)$$

where C_{xy} and C_{yx} are the covariances between variables x and y where as C_{xx}, C_{yy} are auto covariances of variables x and y respectively. There are various applications for CCA in the signal processing domain. It has been useful in finding relations which can help for multi-view learning (Liu et al., 2018). Heycem et.al. applied the technique for feature selection for the problem of depression recognition from speech signals (Kaya et al., 2014). Wang et.al. used CCA to learn acoustic features that can improve phonetic recognition (Wang et al., 2015). Apart from the above mentioned applications, CCA is also useful in areas like Blind Source Separation (BSS). The problem aims to recover the original signal when an unknown linear mixture of statistically independent signals are available (Borga and Knutsson, 2001). Another approach based on CCA focuses to improve the signal to noise ratio (SNR) in EEG data that is recorded from multiple channels (de Cheveigné et al., 2019).

In the present work, knowledge from a set of multiple features is used to detect boundary points in a word. The complete procedure is explained in Section 4..

4. Proposed Approach for Segmentation

The proposed method uses cumulative knowledge of multiple geometric features and use that to form a multi-view trajectory feature vector. The feature vector is then analyzed dynamically to extract the phonetic boundaries. There are

1. Basic feature set (τ)
2. Derived features (τ_D)
3. Multi-view boundary detection algorithm

Each component is explained in next subsequent subsections. Basic and derived features are defined in the next subsection. The segmentation algorithm is explained in Section 4.2.

4.1. Trajectory Features

A speech signal records the nature of vibrations when the vocal chord moves for uttering a sound. The resultant waveform consists of peaks and valleys which helps to understand salient features of the spoken unit and person who has uttered. Thus the waveform records different acoustic events which can be used for various purposes like classification, segmentation, etc. One of the crucial properties of a trajectory is its shape. Each event that is recorded in a speech signal can be distinct in structure. The structural properties of phonetic units have become an interesting area of study (Minematsu, 2005). The reason for this is that the features corresponds to phonetic characteristics with variations in a lucid way. And also the structural properties of waveform trajectories are useful in understanding the dynamic nature of different phonetic units. In the present work, a set of geometric features are proposed to capture the transitional behavior of the waveform that can be further used in identifying boundary points between different phonetic units. The feature set as a whole contains two different classes i.e. primitive and derived properties. The primitive properties are those characteristics that are inherent in a waveform. They are listed as follows:

1. Peak
2. Valley
3. Peak position

4. Valley position

In the second stage, the aforementioned features are transformed further to obtain derived attributes. This set contains the following elements:

1. Peak width
2. Valley width
3. Slope of peaks and valleys
4. Disparity of peaks and valleys

For a segment of speech signal $S[n]$ with size m, the terms are defined in Definitions 1 to 8.

Definition 1 A data point p_i is said to be as **peak** if $p_{i-1} < p_i > p_{i+1}$ where $\forall i \in \mathbb{Z}$

Definition 2 A data point p_i is said to be a **valley** if $p_{i-1} > p_i < p_{i+1}$ where $\forall i \in \mathbb{Z}$

Definition 3 **Peak position** is any integer k, such that $0 < k < m$ where peak is found at k^{th} location

Definition 4 **Valley position** is any integer k, such that $0 < k < m$ where valley is found at k^{th} location

Definition 5 The data point p_k being a peak point between the valleys v_q and v_r, the difference $r - q$ is defined as **peak width** for the peak p_k $\forall k, q, r \in \mathbb{Z}$ and $q < k < r$

Definition 6 The data point v_k being a valley point between two peaks p_q and p_r, the difference $r - q$ is defined as **Valley width** of valley v_k $\forall k, q, r \in \mathbb{Z}$ and $q < k < r$

Definition 7 The **slope** between two points $x = (x_1, y_1)$ and $y = (x_2, y_2)$ is defined by Equation 3.

$$Slope(x, y) = \frac{y_2 - y_1}{x_2 - x_1} \tag{3}$$

Definition 8 The **Disparity** between two points p_i and p_k is given by Equation 4.

$$Disparity(p_i, p_k) = \sqrt{(p_i - p_k)^2}, \forall i, k \in \mathbb{Z} \tag{4}$$

To understand the terms, let us consider Figure 1. In the figure, peaks and valleys are indicated as P_i and V_i respectively where i represents the sequence in which they occur in a waveform. The next term, peak-width is the width of the curve in a waveform between two valley positions. In the same way, valley width is the distance between two peaks in which a valley is present. Slope is the general gradient between two points in a geometric space. The points that are considered here are a pair of peaks (or valleys). This feature gives information of two adjacent peaks (or valleys). In the segmentation algorithm, the average slope between peaks (and valleys) of each frame in the source signal is studied. Finally, the property 'Disparity' between two points (peaks or valleys) is the continuous variation between the heights of peaks and depth of valleys. The property 'slope' considers the position at which the peaks (or valleys) occur whereas 'Disparity' does not regard this property. The derived features of the word "Zero" are shown in Figure 2. Figure 2-a shows the normalized source signal, Figure 2-b and Figure 2-c give slope and disparity of peaks respectively. Slope and disparity of valleys are shown in Figure 2-d and Figure 2-e respectively. The procedure used for segmentation is explained in next subsection.

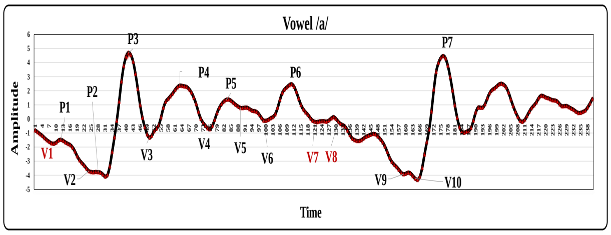

Figure 1: Peaks and valleys of a speech segment

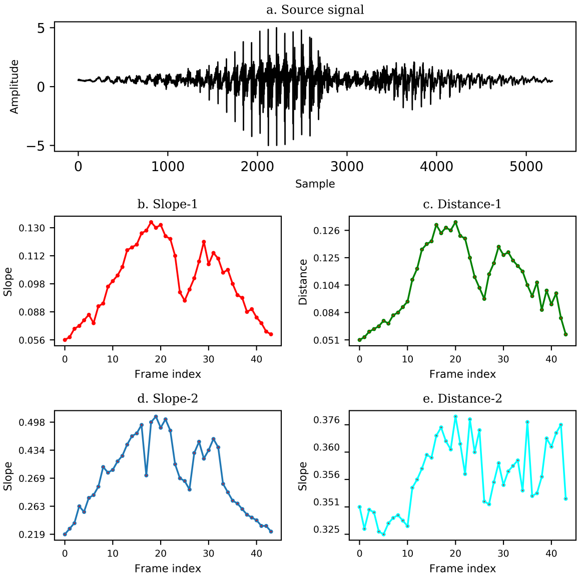

Figure 2: Peak attributes for the word "zero"

4.2. Multi-View Boundary Detection Algorithm

The features that are described in the previous section are analyzed to understand the boundaries of the phonetic units. The algorithm observes the dynamic changes of the waveform over the entire signal by capturing the variations with the extracted features. First, the given speech signal is divided into equal-sized frames and a set of basic features (τ) are extracted from each signal. From the basic features, a set of derived features are drawn. Thus the complete feature set is a matrix in which each set of derived features are present. This is a multi-view representation of the waveform trajectory features that will be processed to find the segmentation points.

The segmentation procedure comprises of two stages: In the first stage, the feature matrix is analyzed by the CCA procedure which will give a set of coefficients for each feature set simultaneously. These coefficients represent the correlation between the subsets of each feature set which will be used next. In the second stage, a pair of sequential frames that are adjacent will be used to generate correlation coefficients. Finally, the coefficients generated in first and second stages are then compared to get the variance between them. The crucial steps in the segmentation procedure can be summarized as follows:

1. The input signal $S[n]$ is divided into a set of frames $f_0, f_1, ..., f_n$ of equal size.
2. Each frame is then transformed to a set of primitive features : S_p, S_v, S_{pi}, V_{vi}, where:
   - S_p is set of peaks
   - S_v is set of valleys
   - S_{pi} is set of integers that represent peak positions
   - V_{vi} is set of integers that represent valley positions
3. The features obtained in Step 2 are then transformed to a set of trajectory features $\tau = \tau_{sv}, \tau_{sp}, \tau_{dpv}, \tau_{dp}$.
4. The feature sets τ are analyzed using CCA which gives a set of coefficients represented by CCA_τ.
5. The features sets belonging to subsequent frames are correlated to get the new coefficients. Each set consists of features belonging to 3 adjacent frames. The number of frames is empirically chosen so that variations can be captured in the corresponding CCA coefficients.
6. Variance between coefficients computed in Step 4 and Step 5 are compared. The peaks in this set forms the boundary points. Thus the peaks in each set are combined to identify the boundary points using the CCA_τ computed by Equation 5.

$$B_p = \{CCA_{\tau_{dp}} \cup CCA_{\tau_{dpv}} \cup CCA_{\tau_{sp}} \cup CCA_{\tau_{sv}}\} \qquad (5)$$

The final variances obtained for each derived feature set are shown in Figure 3. From the diagram, it can be observed that the changes needed for identifying the phonemic variations are recorded in as peak points in the final variances. But different varieties of variations can be seen separately from features. Therefore it is required to combine the points obtained from each features to get the final boundary points. The detailed algorithm and the flowchart are given in Algorithm 1 and Figure 4 respectively. In the next section, the background setup used for the experiments is described.

5. Experimental Setup

The algorithms were implemented using Python platform. The CCA implementation that is available in Pyrcca (Bilenko and Gallant, 2016) library was used in the algorithm. The data used in present work is English digits belong to the Indian accent. The speakers belong to different regions (states) in India. They include male and female speakers. We used 50 speakers data in the analysis. Each English digit was recorded 15 times for all speakers. The digits were recorded using the Cool Edit software with 16KHz sampling rate, mono channel and 16 bits resolution. The behaviour of the algorithm for different cases are discussed in the next section.

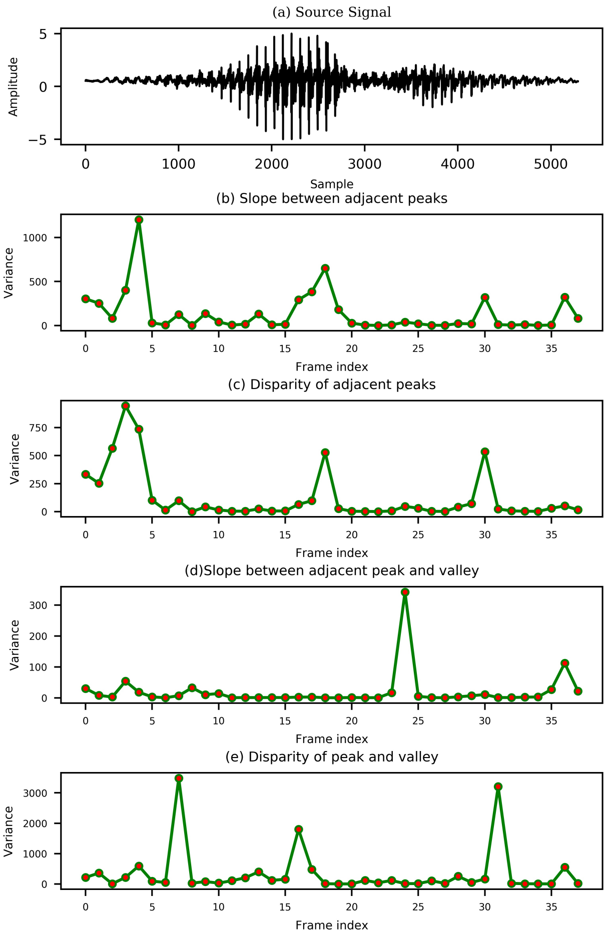

Figure 3: CCA of different features for the word ”zero”

6. Results and Analysis

In the present study, a set of trajectory features are considered to be useful after conducting experiments on various properties. The properties that were observed are shown in Table 1. Figure 5 gives an idea of the nature of these features. They were not used as part of feature set in the segmentation process rather they are useful in understanding the characteristics of regions belonging to different phonetic units. Some observations are presented in each subsequent subsections separately. The analysis of the algorithm's nature for peaks and valleys are presented separately in subsequent subsections.

6.1. Peak Attributes Analysis

To understand meaningful cues from speech, an analysis of the nature of peaks in different classes of sounds like vowels, fricatives and stops are done. These clues are further used to find the boundaries of phonemes. It is helpful to know the regions where changes are occurring corresponding to the behaviour of attributes. Peaks can be classified into different types based on height and width. Vowels like /i/ and /e/ have the regions with higher peaks and vowels /a/, /o/ and /u/ have wider peaks. Figure 5 shows different statistics of peaks. We can understand that the vowel regions have comparatively more wider peaks than non-vowel regions. The analysis of slope was carried in two ways:

Algorithm 1: Boundary_detection algorithm

```
Input:
S[n]: Speech segment of length $n$
k: Size of the frame
Output:
BP: Boundary points of phonetic units
begin
    Step 1: Normalize S[n]
    Step 2: Divide S[n] into frames with equal size $k$
    Step 3: Let $F_n$ be number of frames
    for $i \leftarrow 0$ to $F_n$ do
        Step 3.1: Find peaks using Definition 1
        Step 3.2: Find valleys using Definition 2
    Step 4: for $i \leftarrow 0$ to $F_n$ do
        for $j \leftarrow 0$ to $Max(n_{peaks}, n_{valleys})$ do
            Step 4.1
            $T_{sp} \leftarrow Slope(peaks_j, peaks_{j+1})$
            Step 4.2
             $T_{sv} \leftarrow Slope(valleys_j, valleys_{j+1})$
            Step 4.3
             $T_{dp} \leftarrow Disparity(peaks_j, peaks_{j+1})$
            Step 4.4
             $T_{dv} \leftarrow Disparity(valleys_j, valleys_{j+1})$
        $\tau_i \leftarrow \{T_{sp_i}, T_{sv_i}, T_{dp_i}, T_{dv_i}\}$
    Step 5:
    for $i \leftarrow 0$ to $F_n$ do
        canonicalcoef$_i \leftarrow CCA(\tau_i)$
    Step 6:
    for $i \leftarrow 0$ to $F_n$ do
        coeffnew$_i \leftarrow$
          $CCA_{validate}((\tau_i, ..., \tau_{i+3}), (\tau_{i+3}, ..., \tau_{i+6}))$
        $variance_i \leftarrow$
          $CCA_{Variance}($canonicalcoef$_i$, coeffnew$_i)$
    Step 7: BP$\leftarrow$
      $peaks(variance_{sp}) \cup peaks(variance_{sv}) \cup$
      $peaks(variance_{dp}) \cup peaks(variance_{dv})$
    return BP
```

S.No.	Attribute
1	Peak
2	Peak width
3	Peak position
4	Average difference between adjacent peak values
5	Average slope between adjacent peak values
6	Valley
7	Valley width
8	Valley position
9	Average difference between adjacent valley values
10	Average slope between adjacent valley values

Table 1: Attributes used for analysis

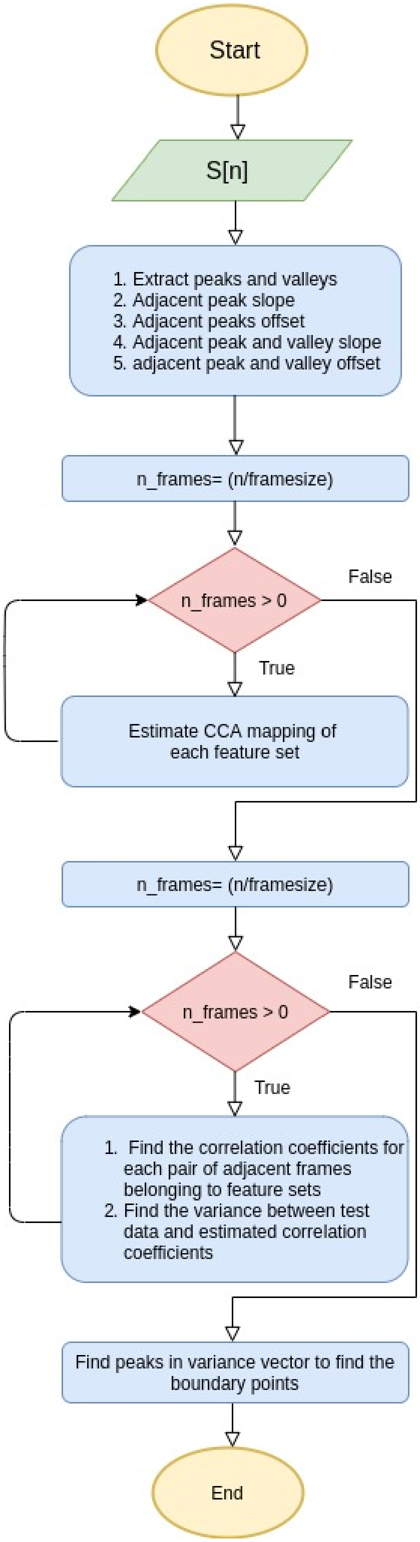
Start
S[n]
1. Extract peaks and valleys
2. Adjacent peak slope
3. Adjacent peaks offset
4. Adjacent peak and valley slope
5. adjacent peak and valley offset
n_frames= (n/framesize)
n_frames > 0
False
True
Estimate CCA mapping of each feature set
n_frames= (n/framesize)
n_frames > 0
False
True
1. Find the correlation coefficients for each pair of adjacent frames belonging to feature sets
2. Find the variance between test data and estimated correlation coefficients
Find peaks in variance vector to find the boundary points
End

Figure 4: Flowchart for the boundary detection algorithm

1. Slope between adjacent peaks in the same frame
2. Slope between peaks of adjacent frames

This attribute is used for understanding structural significance at phoneme boundaries. Slope between adjacent peaks in the same frame does not have much variations.The difference between frames belonging to the same phonetic unit is small. But it is observed that this value is more at the phoneme boundaries. Slope between peaks of vowel regions and non-vowel regions give enough variations that helps in understanding the boundary points. Figure 6 and Figure 7 show slope and disparity between peaks of adjacent frames for the words ”Zero” to ”Nine”. It can be observed that the changes in the wave forms are evident so that structural clues can be captured by features. There has been

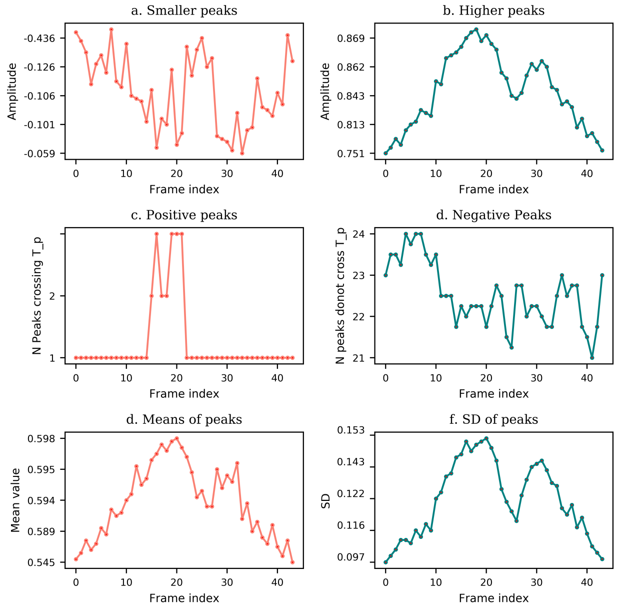

Figure 5: Peak statistics of the word ”zero”

an interesting phenomena observed especially in vowel regions. There is a linear growth of the slope and disparity at the beginning of the vowel region and they start decaying at the middle part and continuing till the boundary is reached. This nature is observed both in intra-frame and inter-frame situations. There is a sudden increase in the slope value at the boundaries of different phonemes.

The average disparity between peaks within vowel region is more than non-vowel regions. Figure 7 shows the disparity between peaks for the word ”Zero”. We can observe that there are prominent changes at boundary frames. The distance between inter frame analysis is to understand the nature of the peak values with their neighbouring frames. This distance is more at the phoneme boundaries when compared to interior regions of phonemes. Anyhow this value is high in vowel regions similar to intra-frame difference. The difference between two frames is stable in the regions belonging to the same phoneme. Therefore it is inferred that intra-frame difference can be used to identify the syllable boundaries whereas inter frame difference is useful in identifying phoneme boundaries. Figure 12 shows distance between peaks in adjacent frames for the word ”Zero”. It also shows that changes can be observed clearly at boundary frames of phoneme or syllable.

6.2. Valley Attributes Analysis

The second crucial feature of waveform in the framework is valley attributes. In this class, the nature of valley was studied by understanding the properties of deeper valleys, higher valleys, positive valleys, negative valleys, etc. Figure 10 shows the statistics of these attributes. The mean and standard deviation of these properties of valleys are shown in each sub figure. These graphs suggest that there is a temporal variation across the frames in these statistics which implies that the properties are significant for phoneme boundary analysis. We can understand variations in valleys for different segments of the speech sub-units. Useful observations from the analysis are listed below:

1. Deeper valleys and shallow over valleys are found more in vowel regions than non-vowel regions.
2. Valleys in vowels are wide.
3. Standard deviation in vowel regions are comparatively higher than non-vowel regions.

These qualities mean that the structural variation can be achieved from valley features also. For example, vowels /i/ and /o/ have differences in the properties in terms of valleys. Vowel /i/ has deeper valleys compared to vowel /o/. It shows that there is more deviation between vowel and non-vowel regions. These statistics suggest that it is meaningful to use valley properties for understanding structural significance. The two properties Slope and disparity of the words ”Zero” to ”Nine” are shown in Figure 8 and Figure 9 respectively. We can see the structural consistency in different utterances of the same digit for a speaker.

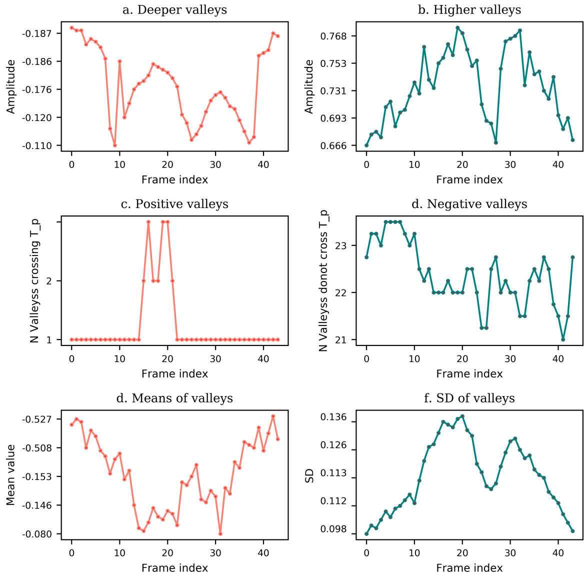

Figure 6: Valley statistics for the word ”zero”

6.3. Characteristics of Method in Noisy Conditions

The method was also evaluated in the presence of noise in input signals. Here, the white noise up to 20dB SNR was considered. Figure 11 shows a source speech signal along with the CCA coefficients of each feature vector. A comparison between Figure 3 and Figure 11 helps in understanding the nature of the algorithm in noisy signals. The first point to understand is that there is a variation in structure of same feature vectors. In this example, the disparity vector differs in variance of CCA coefficients. The noise presence makes the adjacent frames belonging to two different phonetic units much higher in their variation that is reflected in the CCA coefficients. The multi-view analysis enables the method to learn necessary clues from different vectors. Therefore, the failure of capturing the boundary points in one case does not influence much in the final boundary points. So the results suggest that the proposed approach can be effective in noise conditions also.

6.4. Performance of the Algorithm

The proposed approach is successful in identifying the boundary points in 90% of the cases. The mis-identification of boundary points are influenced by speaker's characteristics in failure cases. This include accent, pauses between the phonetic units, etc. The time complexity of the approach includes two major parts including feature extraction step and CCA. Time complexities of different steps are as follows:

1. Peak and valley computation: $\mathcal{O}(n)$.
2. Finding the trajectory properties need constant time $\mathcal{O}(1)$ for each elementary operation which constitutes a linear time complexity $\mathcal{O}(n)$ for n samples.
3. Lastly, CCA algorithm requires $\mathcal{O}(n^3)$ time complexity equivalent to eigen value decomposition method (Uurtio et al., 2017).

Therefore total time complexity of the approach works out to [$\mathcal{O}(n)$ + 4 x $\mathcal{O}(n)$ + 2 x $\mathcal{O}(n^3)$]. The run time requirement of the method is approximately 470 milli seconds. The method was tested on a system with the following configuration:

- Processor : i5 (3.20 GHz)
- Memory : 8 GB

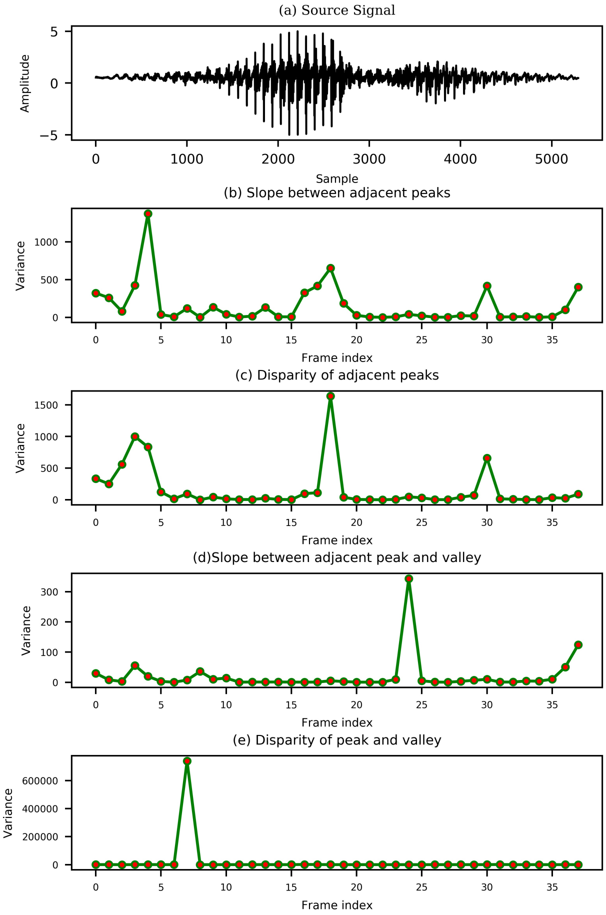

Figure 7: CCA of different features for the word ”zero” (Noisy signal)

7. Conclusions and Future Work

In this paper, a phoneme segmentation approach based on multi-view geometrical features is proposed. The structural properties of speech trajectories are used to find the boundaries between phonetic units using the CCA method. The dissimilarities in geometrical features across a speech trajectory are used as parameters to identify boundary points. To prove the approach, Indian accented spoken English digits data was used in the experiments. The experiments gave reasonable results from which we can infer that the method is effective in identifying the boundary points. Since the approach does not require a training process, the requirement of large data sets are dispensed with. Also as the complexity of the method is reasonable, the run time is less and hence the method is very suitable for low or zero resource languages. The dataset is shared in [1] for the future use of the researchers. The method is being studied at the sentence level for the Hindi language that is spoken in India.

8. Bibliographical References

Ali, A. A., Van der Spiegel, J., Mueller, P., Haentjens, G., and Berman, J. (1999). An acoustic-phonetic feature-based system for automatic phoneme recognition in continuous speech. In *Circuits and Systems, 1999. ISCAS'99. Proceedings of the 1999 IEEE International Symposium on*, volume 3, pages 118–121. IEEE.

Atev, S., Miller, G., and Papanikolopoulos, N. P. (2010). Clustering of vehicle trajectories. *IEEE Transactions on Intelligent Transportation Systems*, 11(3):647–657, Sep.

Bilenko, N. Y. and Gallant, J. L. (2016). Pyrcca: regularized kernel canonical correlation analysis in python and its applications to neuroimaging. *Frontiers in neuroinformatics*, 10:49.

Borga, M. and Knutsson, H. (2001). A canonical correlation approach to blind source separation. *Report LiU-IMT-EX-0062 Department of Biomedical Engineering, Linkping University*.

de Cheveigné, A., Liberto, G. M. D., Arzounian, D., Wong, D. D., Hjortkjær, J., Fuglsang, S., and Parra, L. C. (2019). Multiway canonical correlation analysis of brain data. *NeuroImage*, 186:728 – 740.

Deng, L. and Strik, H. (2007). Structure-based and template-based automatic speech recognition - comparing parametric and non-parametric approaches. In *INTERSPEECH*.

Jeung, H., Shen, H. T., and Zhou, X. (2008). Convoy queries in spatio-temporal databases. In *2008 IEEE 24th International Conference on Data Engineering*, pages 1457–1459, April.

Kaya, H., Eyben, F., Salah, A. A., and Schuller, B. (2014). Cca based feature selection with application to continuous depression recognition from acoustic speech features. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3729–3733. IEEE.

Li, H., Liu, J., Wu, K., Yang, Z., Liu, R. W., and Xiong, N. (2018). Spatio-temporal vessel trajectory cluster-

[1] IITG DIGITS: https://drive.google.com/drive/folders/1px1p2p5QRNNvFvLJT9hgkA93N7$_{U}twzs$

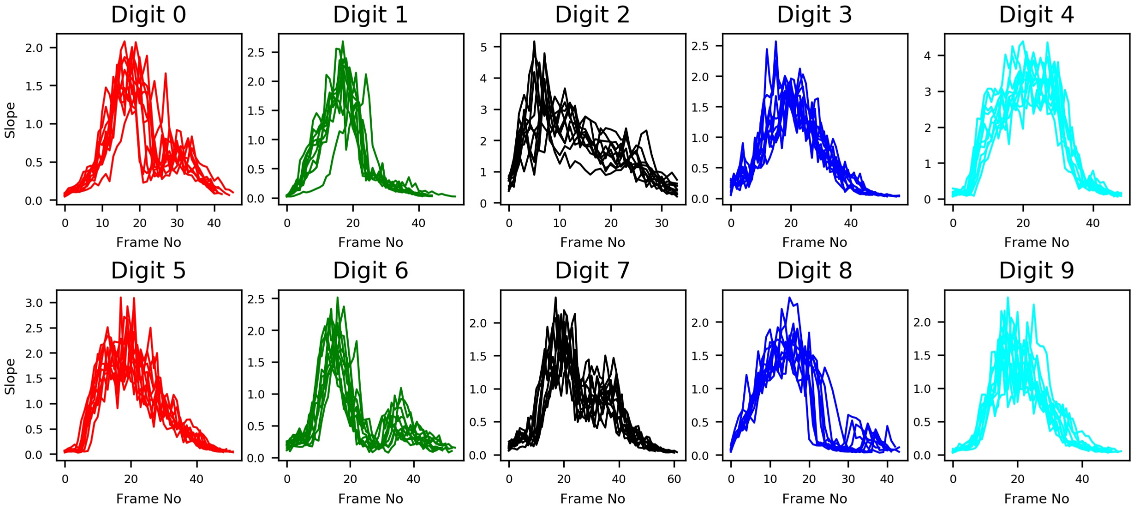

Figure 8: Slope between peaks of the words ”Zero” to ”Nine” for a speaker

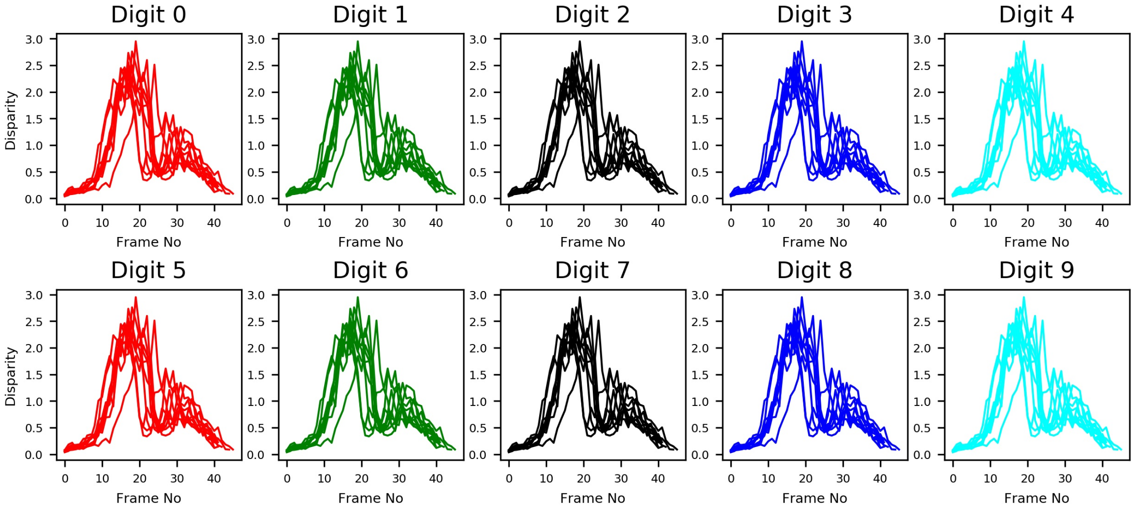

Figure 9: Disparity between peaks of the words ”Zero” to ”Nine” for a speaker

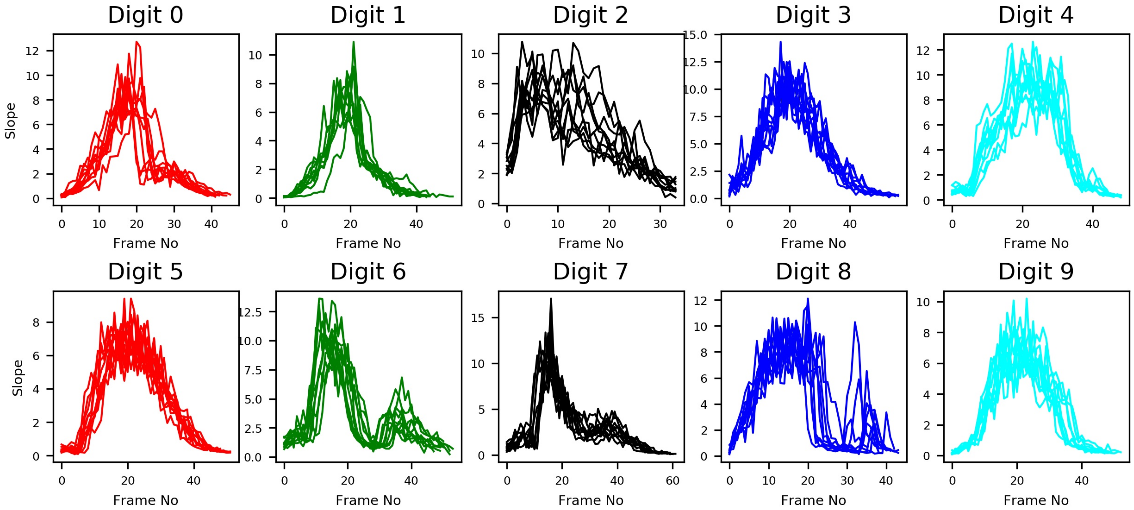

Figure 10: Slope between valleys of the words ”Zero” to ”Nine” for a speaker

ing based on data mapping and density. *IEEE Access*, 6:58939–58954.

Lin, Z., Zeng, Q., Duan, H., Liu, C., and Lu, F. (2019). A semantic user distance metric using gps trajectory data. *IEEE Access*, 7:30185–30196.

Liu, S. and Sim, K. C. (2012). Implicit trajectory modelling using temporally varying weight regression for automatic speech recognition. In *2012 IEEE International*

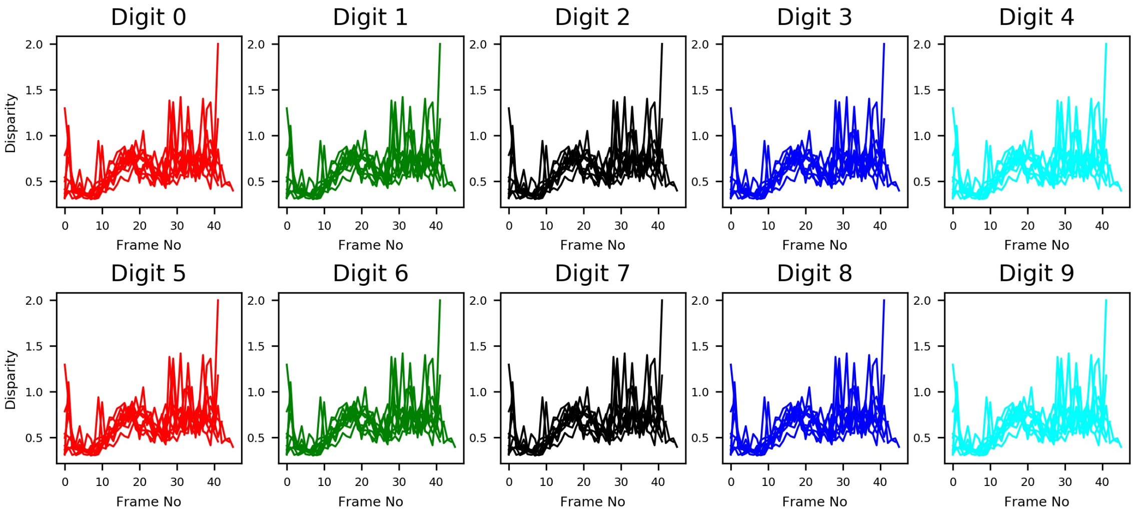

Figure 11: Disparity between valleys of the words ”Zero” to ”Nine” for a speaker

Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4761–4764, March.

Liu, Y., Li, Y., and Yuan, Y.-H. (2018). A complete canonical correlation analysis for multiview learning. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3254–3258. IEEE.

Minematsu, N. (2005). Mathematical evidence of the acoustic universal structure in speech. In *Proceedings. (ICASSP ’05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I/889–I/892 Vol. 1, March.

Mitra, V., Wang, W., Stolcke, A., Nam, H., Richey, C., Yuan, J., and Liberman, M. (2013). Articulatory trajectories for large-vocabulary speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7145–7149, May.

Russell, M. J. and Holmes, W. J. (1997). Linear trajectory segmental hmms. *IEEE Signal Processing Letters*, 4(3):72–74, March.

Siohan, O. and Yifan Gong. (1996). A semi-continuous stochastic trajectory model for phoneme-based continuous speech recognition. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 471–474 vol. 1, May.

Uurtio, V., Monteiro, J. M., Kandola, J., Shawe-Taylor, J., Fernandez-Reyes, D., and Rousu, J. (2017). A tutorial on canonical correlation methods. *ACM Computing Surveys (CSUR)*, 50(6):1–33.

Wang, W., Arora, R., Livescu, K., and Bilmes, J. A. (2015). Unsupervised learning of acoustic features via deep canonical correlation analysis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4590–4594, April.

Xiao, P., Ang, M., Jiawei, Z., and Lei, W. (2019). Approximate similarity measurements on multi-attributes trajectories data. *IEEE Access*, 7:10905–10915.

Yan, R., Zu, Y., and Zhu, Y. (2006). Automatic speech segmentation combining an hmm-based approach and recurrence trend analysis. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages I–I. IEEE.

Yifan Gong. (1997). Stochastic trajectory modeling and sentence searching for continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 5(1):33–44, Jan.