Characterizing Speaker Variability Using Spectral Envelopes of Vowel Sounds

A. N. Harish, D. R. Sanand and S. Umesh

Department of Electrical Engineering, Indian Institute of Technology-Kanpur, Kanpur, India-208016

{harishan, drsanand, sumesh}@iitk.ac.in

Abstract

In this paper, we present a study to understand the relation among spectra of speakers enunciating the same sound and investigate the issue of uniform versus non-uniform scaling. There is a lot of interest in understanding this relation as speaker variability is a major source of concern in many applications including Automatic Speech Recognition (ASR). Using dynamic programming, we find mapping relations between smoothed spectral envelopes of speakers enunciating the same sound and show that these relations are not linear but have a consistent non-uniform behavior. This non-uniform behavior is also shown to vary across vowels. Through a series of experiments, we show that using the observed non-uniform relation provides better vowel normalization than just a simple linear scaling relation. All results in this paper are based on vowel data from TIMIT, Hillenbrand et al. and North Texas databases.

Index Terms: vowel normalization, non-uniform scaling, dynamic programming, speaker variability

1. Introduction

Speaker variability has been a widely researched topic for many years because of its significant influence on the performance of Speaker Independent Automatic Speech Recognition (SI-ASR) systems. Differences in Vocal Tract Length (VTL) are known to be responsible for variations in spectra of the same sound enunciated by several speakers. Nordström and Lindblom [1] proposed a normalization scheme in which all formants of a speaker are scaled by a single scale factor based on the estimate of the speaker’s vocal tract length. Fant [2] argued that this was not appropriate and proposed normalization by a scaling factor that depends on the vowel category and formant number. Other methods proposed for normalization include log-interval normalization used by Nearey which is based on the mean log value of speakers’ first two formants [3]. Syrdal and Gopal [4] came up with a scheme based on the theory that vowels are relative patterns and not absolute formant frequencies. Umesh et al. [5] have proposed a warping function that depends only on frequency and has a performance equal to that of Fant’s method. They have also proposed an affine transformation that relates formant frequencies of speakers [6]. However, the issue of uniform versus non-uniform scaling remains unresolved and is the subject of this study.

The approaches mentioned above make use of formant frequencies of vowels to perform normalization. In this paper we explore the relationship among male, female and child speakers by using smoothed spectral envelopes of vowel sounds. Since most ASR systems use spectral based features, we seek to reduce variability in the spectra of speakers enunciating the same sound and in turn normalize their features. Hence our approach examines spectral envelopes rather than formant frequencies. Using dynamic programming techniques, we show that speakers are not related by a simple linear function of frequency. We also corroborate our observations with the help of a series of experiments performed on three vowel databases.

The paper is organized as follows. Section 2 describes our approach for deriving relations between spectra of speakers using an algorithm based on dynamic programming. In section 3, we analyze the behavior of mapping relations obtained from our algorithm. In section 4, we present results of normalization experiments performed using our derived mapping relations and compare them with the uniform scaling approach. Section 5 concludes the paper and discusses the scope for future work.

2. Dynamic Programming (DP) Approach for Analysis of Speaker Variability

2.1. Obtaining smoothed spectra

Smoothed spectra are obtained on 20ms frames extracted from the steady state portion of vowel utterances. We follow the conventional filter-bank smoothing technique used in Mel Frequency Cepstral Coefficient (MFCC) feature extraction [7]. We have chosen a filter-bank with 256 channels so that the resulting envelope is sufficiently smooth and continuous. The bandwidth of a channel is chosen such that it includes a sufficiently large number of spectral samples for smoothing.

2.2. Finding a reference spectrum

We make use of formant information in order to find a reference female and a reference male spectrum for every vowel sound in a database. If formant data are not provided in the database, we estimate them using Linear Prediction Coefficients (LPCs) since reference spectrum calculation does not demand the use of highly accurate formant estimates.

In order to find the reference spectrum for a vowel the following operations are performed

- The means and standard deviations of the first three formants ($F_1$, $F_2$ and $F_3$) over all speakers of a particular gender (male or female) are computed for that vowel.
- The spectra of all speakers of that gender whose $F_1$, $F_2$ and $F_3$ lie within 20% of the standard deviation on either side of their corresponding means are averaged to obtain the required reference spectrum.

Fig. 1 shows the reference male and female speaker spectra obtained for the vowel /uw/ in Hillenbrand data. The term reference speaker is used interchangeably with reference spectrum in rest of the paper.
Figure 1: Female and male reference spectra for the vowel /uw/ in Hillenbrand data.

2.3. Details of dynamic programming based algorithm

The approach followed in our algorithm is similar to that of dynamic time warping [8]. Instead of a time sequence of values, our test and template patterns here are the test and reference speaker formants. Let the smoothed spectra of the reference and test speakers be denoted by $R$ and $T$ respectively. Slope vectors $s_{t,y}$ at the $y^{th}$ point of $T$ and $s_{r,x}$ at the $x^{th}$ point of $R$ are calculated as

\[ s_{t,y} = [T(y-2) - T(y-3), T(y-1) - T(y-2), T(y) - T(y-1), T(y+1) - T(y), T(y+2) - T(y+1)] \]

\[ s_{r,x} = [R(x-2) - R(x-3), R(x-1) - R(x-2), R(x) - R(x-1), \hat{R}(x+1) - \hat{R}(x), \hat{R}(x+2) - \hat{R}(x+1)] \]

(1)\hspace{1cm} (2)

A square grid is formed with its size equal to the length of smoothed spectra. The value associated with point $(x, y)$ in the grid is calculated as

\[ d(x, y) = ||s_{r,x} - s_{t,y}|| \quad \text{(3)} \]

Note that the reference and test spectral envelopes are both of length $L$. Starting at the point $(x, y) = (1, 1)$ in the grid, the algorithm performs the following operation recursively until $(x, y) = (1, 1)$.

\[ (x^*, y^*) = \arg \min_{k, i, j} (d(x, y) + d(x - i, y - j)) \quad \text{(4)} \]

The sequence of $(x^*, y^*)$ pairs from $(1, L)$ to $(1, 1)$ obtained from (4) determine the optimal path through the grid.

We impose a set of local constraints on the path that can be taken through the grid by allowing transitions to any given point $(x, y)$ from just three neighboring points $(x - 1, y - 1)$, $(x - 1, y)$ and $(x, y - 1)$. Each transition is associated with a weight which is denoted as $k$ in (4). The diagonal transition is associated with a weight of 2 and the other two transitions with a weight of 1. We also confine the slope of the grid to which the best path is searched to a portion of the grid depending on the relation between test and reference speaker formants.

3. Analysis of Mapping Relations Obtained from DP Based Approach

In our experiments, we make use of vowel utterances from the following databases

- TIMIT vowels extracted from the standard TIMIT database, having 13 vowels with 150 male and 150 female utterances per vowel.
- North Texas vowel database [10] having 12 vowels with 75 male and 75 female utterances per vowel.

Figure 2: Mapping relations obtained for all vowels in North Texas data illustrating the deviation from origin.

3.1. Finding mappings for the average speaker

Using the procedure outlined in sec. 2, a mapping relation is found between a speaker’s spectrum for a vowel and the reference speaker for that particular vowel. Two mappings are found for every speaker, one with respect to the reference female and one with respect to the reference male. This is done for all speakers and vowels in the database. The mappings for all speakers of a particular category (male, female or child) with respect to a particular reference are found to behave in a similar fashion. Hence we obtain two representative relations (with respect to reference male and female) for a category by averaging the mappings between speakers in that category and the particular reference speaker. These representative relations are henceforth referred to as mappings for the average speaker (male, female or child).

Fig. 2 shows mappings between the average male and reference female and vice-versa for all vowels in North Texas data. Fig. 3 shows mappings between the average child and reference female for vowels /ae/ and /uh/ in Hillenbrand data. In Fig. 4(a), mappings between average male and reference female and vice-versa are plotted for the vowel /ae/ of all three databases. They are observed to be similar for the three databases. The ratio of reference frequency to the corresponding mapped frequency (scaling function) for the plots in Fig. 4(a) is plotted in Fig. 4(b).

3.2. Observations

Our experiments lead us to the following observations

- As observed in Fig. 2, the mapping relations do not start increasing from the origin in a linear fashion but show a small deviation from the origin before increasing monotonically. This behavior is observed in all vowels in all three databases.
- The mapping relations change slope at least once after the
first formant frequency as demonstrated with the help of plots in Fig. 3 and Fig. 4(a). However, the frequency at which this slope change occurs varies across vowels. As seen from Fig. 3, the mapping relations for /ae/ and /ah/ change slope at different frequencies. This behavior is prominent in relations between speakers of different categories, like between a test male and reference female and between a test child and reference male. Sometimes, more than one slope change may be observed in a mapping relation.

As demonstrated by the scaling functions in Fig. 4(b), the scaling factor clearly depends on frequency as opposed to linear scaling where it takes a constant value independent of frequency. Here, the scaling factor becomes fairly constant after $F_3$ and settles at a value that depends on the speaker. This is in agreement with the fact that for a particular enunciation, higher order formants of different speakers are approximately linearly related according to their vocal tract lengths [11].

From the above observations it is clear that a non-uniform (or non-linear) relation exists among the spectra of speakers enunciating the same sound.

4. Vowel Normalization Results

Here we show normalization results obtained by using the scaling functions derived from our approach and compare them with the uniform scaling. All results in sec. 4.1 and 4.2 are arrived at by considering female as reference.

4.1. Formant normalization

Here we show formant normalization results for vowels in Hillenbrand data since the database provides carefully estimated formant frequencies for every vowel utterance. Fig. 5 shows $F_1 - F_2$ and $F_2 - F_3$ plane normalization of the back vowels /aw/, /oa/, /oo/ and /ah/ in Hillenbrand data. It is clear from the figure that non-uniform scaling not only reduces the standard deviation of formants marked along their respective axes.

4.2. Feature space normalization

Here we compare the normalization performance of uniform and non-uniform scaling methods with the help of class separability in the feature space. We make use of the first 13 values of MFCC feature vectors obtained by taking a Discrete Cosine Transform (DCT) of smoothed spectra.

Since discriminability in the feature space is important from the point of view of ASR systems, a good measure of the normalization performance would be the F-Ratio. Let $M_i$ and $R_i$ denote the mean MFCC vector and its covariance matrix, respectively, for the $i^{th}$ vowel. An equal probability of the vowels being compared is assumed. Let $M_0 = \frac{1}{t} \sum_{i=1}^{t} M_i$ where $t$ is the number of vowels being compared. Then the within-class and between-class scatter matrices $S_w$ and $S_b$ are computed as

$$S_w = \frac{1}{t} \sum_{i=1}^{t} R_i \quad S_b = \frac{1}{t} \sum_{i=1}^{t} (M_i - M_0)(M_i - M_0)^T$$  \hspace{1cm} (5)

The separability criterion (F-Ratio) is then given by

$$J = \text{trace}(S_b^{-1}S_w)$$  \hspace{1cm} (6)
Figure 6: Visual representation of pairwise cepstral space vowel separability in TIMIT data. Lighter shades of gray correspond to higher values of separability.

Using the criterion in (6), we compute pairwise separability for all possible vowel pairs in a database. Non-uniform scaling yields higher separability than uniform scaling in all the vowel pairs of TIMIT data, 52 out of 66 pairs in Hillenbrand data and 51 out of 66 pairs in North Texas data. Fig. 6 shows the separability of vowel pairs in TIMIT data pictorially where lighter shades of gray are used to denote higher values of separability.

Using (6), we also compute the overall class separability (i.e. considering all vowels in (5) and (6)) for a database. Table 1 shows the overall separability for the three databases in terms of F-Ratios. Non-uniform scaling shows 26.2%, 7.4% and 20.4% improvement over uniform scaling for TIMIT, Hillenbrand and North Texas data respectively.

4.3. Single gaussian vowel recognition experiment

Here we present results of normalization using statistical models without assuming the knowledge of a reference speaker. The dataset consists of 3900 frames for training and 3625 frames for testing drawn from vowels in TIMIT data. Each vowel is modeled using a single 13 dimensional gaussian having a diagonal covariance matrix. The mean vector and covariance matrix for each vowel are estimated from the feature vectors of train data. This gives us a set of baseline models \( \{ \lambda_i \} \) where \( V \) is the number of vowels.

In case of uniform scaling, unwarped train data for each vowel are warped by choosing a warp factor in the range \([0.65, 1.35]\) that maximizes the likelihood of a warped feature with respect to the corresponding baseline vowel model. The warped feature vectors are used to build normalized models \( \{ \lambda_i^U \} \), where \( L \) indicates linear warping. In order to perform non-uniform scaling, we first obtain 5 mappings for every category of speakers with respect to a particular reference (for every vowel). These 5 mappings are obtained by choosing 5 equally spaced functions from the mappings for all speakers in that category so that the range of speaker variation is covered. These 5 mappings between every category and reference are combined to form a pool of functions for that vowel. Unwarped train data are warped by choosing a function from the pool that maximizes the likelihood of a warped feature with respect to the baseline model. The warped feature vectors are used to build normalized models \( \{ \lambda_i^N \} \), where \( N \) indicates non-linear warping.

Finally, recognition of test data is performed by doing a maximum likelihood search for the best model. Results of recognition of unwarped test data using baseline models \( \{ \lambda_i \} \), uniformly warped test data using \( \{ \lambda_i^U \} \) and non-uniformly warped test data using \( \{ \lambda_i^N \} \) are shown in table 2. These results show that the experimentally obtained mapping relations perform well not only with a reference speaker spectrum but also with statistical models.

Table 1: Overall F-Ratio in the feature space for TIMIT, Hillenbrand and North Texas vowel data.

<table>
<thead>
<tr>
<th>Database</th>
<th>no scaling</th>
<th>uniform</th>
<th>non-uniform</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIMIT</td>
<td>1.927</td>
<td>2.2291</td>
<td>2.8129</td>
</tr>
<tr>
<td>Hillenbrand</td>
<td>3.4749</td>
<td>3.7328</td>
<td>4.0103</td>
</tr>
<tr>
<td>North Texas</td>
<td>2.5311</td>
<td>3.1004</td>
<td>3.7314</td>
</tr>
</tbody>
</table>

Table 2: Comparison of vowel recognition results between uniform and non-uniform normalization of TIMIT data.

<table>
<thead>
<tr>
<th>Model</th>
<th>% recognition accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>44.2</td>
</tr>
<tr>
<td>uniform normalization (( \lambda^U ))</td>
<td>53.7</td>
</tr>
<tr>
<td>non-uniform normalization (( \lambda^N ))</td>
<td>57.1</td>
</tr>
</tbody>
</table>

5. Conclusion

We have examined the relation that exists among spectra of speakers enunciating the same vowel using dynamic programming and shown that there is a consistent non-linear behavior in the mapping functions that relate them. We have shown using three vowel databases that the mapping relations have a deviation from origin and also change slope after the first formant frequency. With the help of various experiments we have also shown that normalization of formants as well as features is significantly better when the non-uniform behavior is taken into account as compared to a simple linear scaling of spectra. As part of our future work, we wish to explore ways of parameterizing the observed behavior of mapping relations obtained and incorporate them in a practical ASR framework.

6. References