On the Use of Pitch Normalization for Improving Children’s Speech Recognition

Rohit Sinha and Shweta Ghai

Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, Guwahati-781039, India.

Abstract

In this work, we have studied the effect of pitch variations across the speech signals in context of automatic speech recognition. Our initial study done on vowel data indicates that on account of insufficient smoothing of pitch harmonics by the filter bank, particularly for high pitch signals, the variances of mel frequency cepstral coefficients (MFCC) feature significantly increase with increase in the pitch of the speech signals. Further to reduce the variance of MFCC feature due to varying pitch among speakers, a maximum likelihood based explicit pitch normalization method has been explored. On connected digit recognition task, with pitch normalization a relative improvement of 15% is obtained over baseline for children’s speech (higher pitch) on adults’ speech (lower pitch) trained models.

Index Terms: children’s speech recognition, pitch normalization, mel frequency cepstral coefficients

1. Introduction

The acoustic and the linguistic attributes such as pitch, formant frequencies, average phone duration, speaking rate, pronunciation and grammar in case of children have been reported to differ largely from those of the adults [1][2]. On account of these variations, the recognition performance of automatic speech recognition (ASR) systems trained on adults’ speech degrades significantly for children’s speech [3][4][5]. Among these variations, the role of pitch variations is not well explored. In literature, a few studies have reported improved classification performance with pitch-dependent normalization of cepstra [6] and improved children ASR performance on reduction of pitch of the signals [7]. But as the ASR systems typically use mel frequency cepstral coefficients (MFCC) feature that do not extract the pitch information rather smooth it out, thus reducing the speaker-dependence, the ASR system performance should be rather insensitive to pitch variations among speakers.

Motivated by these facts, in this work, we have studied the effect of pitch on MFCC feature. Our study indicates the existence of a systematic variation in the variances of MFCC due to varying pitch across the speech signals. Further, on exploring the smoothed spectral envelope corresponding to MFCC feature, some pitch-dependent distortions are noticed, in particular for high pitch signals. For reducing the effect of pitch variations among speakers, a pitch normalization scheme, similar to vocal tract length based speaker normalization [8], is explored. The pitch synchronous time scaling (PSTS) method [9] is used for transforming the pitch of the speech signals. The proposed pitch normalization scheme results in significant improvement in the recognition performance of children’s speech whose pitch values are usually much higher than those of the adults’ speech.

The rest of the paper is organized as follows. Section 2 presents the analysis of the effect of pitch variations on MFCC feature. Section 3 describes the experimental setup and the database used in this work. Section 4 presents the experimental results followed by conclusion in Section 5.

2. Effect of Pitch Variation on MFCC

In this section, we present a study performed using the TIMIT database to illustrate the effect of pitch variations on MFCC feature. To be consistent with the recognition experiments, reported later in this work, the vowel speech data from the TIMIT database is downsampled from 16 kHz to 8 kHz.

To begin with, the speech signals from the TIMIT database belonging to low (75-100 Hz) and high (200-250 Hz) pitch ranges are selected. The pitch of the signals is estimated using the ESPS tool available in the Wavesurfer software package [12]. The steady portions of different vowels present in the selected signals are extracted and their corresponding MFCC features are computed. Nearly 2000 frames are used for feature extraction for each vowel.

The plots of mean along with the bar-plots showing variance of each of the 12 dimensions (C_{1} - C_{12}) of MFCC feature for signals belonging to ‘low’ and ‘high’ pitch groups for vowels /æ/, /i:/ and /ə/ are shown in the left panels of Fig. 1(a), (b) and (c), respectively. From these figures, it is noted that the variances of the higher dimensions of the MFCC feature of the high pitch group signals are much larger than those of corresponding to the low pitch group signals. To analyze whether these differences in the variance of MFCC are caused by pitch variations, the pitch of the 200-250 Hz pitch group signals are transformed to 140-175 Hz pitch range through a constant factor of 0.7 for all vowels using the PSTS method. The plots of mean along with the bar-plots showing variance of each of the 12 dimensions of MFCC feature for signals belonging to transformed pitch group and the original pitch group (for sake of comparison) for vowels /æ/, /i:/ and /ə/ are shown in the right panels of Fig. 1(a), (b) and (c), respectively.

From these figures, it is noted that on pitch reduction the variances of the higher dimensions of the MFCC feature reduce considerably. To quantify the effect of these variations in MFCC feature, due to changes in pitch across the signals, on the classification performance on the models trained on features of low pitch group signals (75-100 Hz), the Mahalanobis distance (MD) measure [10] is employed. The MD is computed for cepstral feature vector of all pitch groups using eqn. 1:

$$MD(x, \mu_L) = (x - \mu_L)\Sigma_L^{-1}(x - \mu_L)$$

where $x$ is the cepstral feature vector whose distance from
we further explored the smooth spectra corresponding to MFCC feature for different vowels. The smooth spectrum corresponding to MFCC is derived by computing a 128-point inverse discrete cosine transform of 13 dimensional MFCC feature $(C_1 - C_{12})$. Fig. 2 shows the smooth spectra along with the linear (DFT) spectrum for signals with pitch values of 90 Hz, 220 Hz and 270 Hz for a stable portion of vowel /iy/. Some uncharacteristic distortions are observed in the spectral envelope at the lower frequencies (below 1 kHz) for 220 Hz and 270 Hz pitch signals when compared with that of the 90 Hz signal which enhance with increasing pitch of the signal. We hypothesize that the possible cause of these distortions in the spectral envelope is the insufficient smoothing of the pitch harmonics by the filterbank particularly at low frequencies as the width of the lower order filters is around 100 Hz only. This could explain the cause of the earlier observed increased variances of the higher dimensions of MFCC feature with increasing pitch of the signals.

### 3. Experimental Setup and Databases

The connected digit recognizer used in this work is developed using the HTK toolkit following the setup described in [11]. The 11 digits (0-9 and OH) are modeled as whole word hidden Markov models (HMMs) using 16 states per word. Each state is a mixture of 5 diagonal covariance Gaussian distributions with simple left-to-right moves without any skips over the states. The speech is analyzed with a Hamming window of length 25 ms, frame rate of 100 Hz and pre-emphasis factor of 0.97. Feature vectors of 39 dimensions are used. The feature vectors comprise of $C_0$ to $C_{12}$ static MFCC, their first and second order derivatives. Cepstral mean subtraction is also applied. The word error rate (WER) is used as a measure for recognition performance.

---

Table 1: The variance of the Mahalanobis distances of the original signals of 75-100 Hz and 200-250 Hz pitch groups and the transformed signals with pitch transformation from 200-250 Hz to 140-175 Hz pitch range from the mean of the 75-100 Hz pitch group signals for different vowels.

<table>
<thead>
<tr>
<th>Pitch Range</th>
<th>Vowel /ae/</th>
<th>Vowel /iy/</th>
<th>Vowel /ao/</th>
</tr>
</thead>
<tbody>
<tr>
<td>75-100 Hz (Original)</td>
<td>61</td>
<td>73</td>
<td>39</td>
</tr>
<tr>
<td>200-250 Hz (Original)</td>
<td>2553</td>
<td>1735</td>
<td>1196</td>
</tr>
<tr>
<td>200-250 Hz (Transformed to 140-175 Hz)</td>
<td>283</td>
<td>247</td>
<td>130</td>
</tr>
</tbody>
</table>
The training and the testing data for the digit recognizer have been obtained from the TIDIGITS database. The training set consists 35,566 words from adult male and female speakers having pitch values between 70-250 Hz. The adult test set contains 10,813 words from adults of both sexes having pitch values between 80-260 Hz. The children test set consists of 25,525 words from children of both sexes having age between 6-15 years and pitch values ranging from 100-360 Hz. In this work, all speech data is downsampled to 8 kHz.

4. Experimental Results

The recognition performances (WER) of the adult test set on the connected digit recognition system used in this work is 0.43%. For the children test set the same system gives a WER of 11.37%. The baseline recognition performance of the children test set along with their breakup for different pitch ranges are given in the top row of Table 2.

In order to reduce effect of pitch variations across the adult training and the children test set signals, the pitch of all the children speech signals is first transformed to the mean pitch value of the training set signals i.e., 130 Hz. The recognition performance of children test set with pitch transformed to 130 Hz, along with their breakup for different pitch ranges, is given in the middle row of Table 2. It is to note that since the children test set contains signals with pitch values ranging from 100-360 Hz, transforming pitch of all speech signals to a fixed value of 130 Hz may not be appropriate. This explains the cause of getting performance improvement for only high pitch signals ($F_0 > 300\,Hz$) on transforming pitch of all test signals to a fixed value of 130 Hz. Thus, for determining the appropriate pitch values for the pitch transformed signals, a maximum likelihood (ML)-based grid search is performed.

For doing ML grid search, the pitch of each signal is first transformed to seven different pitch values ranging from 70-250 Hz in a step of 30 Hz. This range of transformed pitch values was chosen based on the fact that the training data has a pitch distribution from 70-250 Hz as shown in the histogram in Fig. 3. Given the various pitch transformed versions within the specified range, the optimal value $\hat{p}$, to which the pitch of each signal is to be transformed to, is estimated as:

$$\hat{p} = \arg \max_{p} Pr(X^p_i|\lambda, W_i)$$  \(2\)

where $X^p_i$ is the feature for the $i^{th}$ utterance with pitch transformed to $p$, $\lambda$ is the speech recognition model and $W_i$ is the transcription of the $i^{th}$ utterance. $W_i$ is determined by initial recognition pass using original feature (i.e., without pitch transformation). The ML search is then performed over the original speech signal and its corresponding seven different pitch transformed versions. The recognition performance of the pitch normalized children speech signals with respect to the baseline model is given in the bottom row of Table 2 along with their breakup for different pitch ranges.

![Figure 2: Plots of smooth spectra corresponding to MFCC feature along with linear (DFT) spectrum for vowel /iy/ of pitch (a) 90 Hz (b) 220 Hz (c) 270 Hz.](image)

![Figure 3: Histogram showing the distribution of pitch of the signals in the training data.](image)

From Table 2, it is noted that the recognition performance of the ML-based pitch normalized children test set has improved by 15% relative over the baseline. On observing the pitch group-wise performances given in Table 2, it is noted that with ML-based pitch normalization consistent improvements are obtained for different pitch groups i.e., higher pitch groups have shown greater improvements. Thus, the performance improvement with explicit pitch normalization, could be attributed to the reduction of the pitch-dependent distortions in the spectral

<table>
<thead>
<tr>
<th>Condition</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>11.37 6.54 17.47 39.03</td>
</tr>
<tr>
<td>Pitch Norm. Fixed Value (To 130 Hz)</td>
<td>13.20 9.49 17.96 33.71</td>
</tr>
<tr>
<td>Pitch Norm. ML-search (70-30-250 Hz)</td>
<td>9.64 6.02 14.24 30.11</td>
</tr>
</tbody>
</table>

Table 2: Performances of the children test set with and without pitch normalization along with breakup for different pitch ranges (The quantity in bracket gives the number of utterances in that pitch group).

![Linear (DFT) vs MFCC](image)
envelope. This can be further understood by observing the histograms of the distribution of pitch of the children test data before and after explicit ML pitch normalization as shown in Fig. 4. It is noted that after explicit pitch normalization, the mode of the pitch distribution of the original children test set signals has shifted towards that of the training data.

5. Conclusion

In this work, the effect of pitch variations across the signals on MFCC feature has been studied. Our preliminary study done on vowel data shows that on account of insufficient smoothing of the pitch harmonics by the filterbank, the spectral envelope corresponding to MFCC feature exhibits pitch-dependent distortions, particularly for high pitch signals. These distortions in the spectral envelope of the speech signals which enhance with increasing pitch of the speech signals are attributed to the increase in the variances of the higher dimensions of the MFCC feature. These increased variances, thus, contribute to the degradation in the recognition performance of the children’s speech data on the adults’ speech trained acoustic models. For reducing the effect of pitch variations on the recognition performance for children speech on the adults’ speech trained models, a maximum likelihood based pitch normalization method has been explored. With ML-based explicit pitch normalization, the children speech recognition performance has been found to improve 15% relative over the baseline for a connected digit recognition task.

In this work, for ML pitch normalization of the speech signals, the pitch of the signals is explicitly transformed to various values in time domain which is a computationally intensive procedure. As we have noted that the improvement with pitch normalization is actually due to reduction in the distortions caused by the insufficient smoothing of the pitch harmonics by the filterbank, we can address this issue more efficiently in the feature domain and this work has been reported in [13].

6. Acknowledgment

Part of this work has been supported by the ongoing UKIERI project titled “Study of Source Features for Speech Synthesis and Speaker Recognition” between IIT Guwahati, IIIT Hyderabad and CSTR, School of Informatics, University of Edinburgh, Edinburgh, UK.

7. References