A Study on the Effect of Pitch on LPCC and PLPC Features for Children's ASR in comparison to MFCC

Shweta Ghai, Rohit Sinha

Department of Electronics and Electrical Engineering,
Indian Institute of Technology Guwahati, Guwahati-781039, India.
{shweta, rsinha}@iitg.ernet.in

Abstract

In this work, following our previous studies, we study and quantify the effect of pitch on LPCC and PLPC features and explore their efficacy for children's mismatched ASR in comparison to MFCC. Our analysis shows that, unlike MFCC, LPCC feature has no major influence of pitch variations. On the other hand, similar to MFCC, though PLPC is also found to be significantly effected by pitch variations but comparatively to a lesser extent. However, after explicit pitch normalization of children's speech, MFCC is found to result in the best children's speech recognition performance on adults' speech trained models in comparison to LPCC and PLPC features.

Index Terms: children's speech recognition, pitch, LPCC, PLPC, MFCC

1. Introduction

In automatic speech recognition (ASR) systems, the objective, given the speech signal, is to find 'what is spoken rather than who has spoken'. Therefore, in the features used for ASR the pitch information is not extracted but rather smoothed out or discarded so as to reduce speaker-dependence. However, some studies in literature indicate that mel frequency cepstral coefficients (MFCC) do capture some pitch-related information [1][2][3]. Also, in [4], it has been reported that MFCC is influenced by the high pitch of the speech signals. Motivated by these, in [5], we explored the effect of pitch variations on MFCC and found significant improvement in the children's ASR performance on the adults' speech trained models after explicit pitch normalization of children's speech.

In [4], the minimum variance distortion less response (MVDR) based spectrum is shown to provide better spectral modeling than DFT-based and LP spectrum especially for medium and high pitch signals. Thus, in [6], we explored the pitch-robustness of the recently proposed MVDR spectrum based perceptual MVDR (PMVDR) [7] feature for children's mismatched ASR in comparison to MFCC but found MFCC to be a better alternative feature for children's mismatched ASR than PMVDR. In literature, few studies have reported that linear prediction cepstral coefficients (LPCC) are also influenced by the high pitch of the speech signals [8]. Also, many studies have reported either comparable or slightly better performance with perceptual linear prediction coefficients (PLPC) than with MFCC for adults' ASR under matched condition [9][4]. Motivated by these, in this work, we study and quantify the effect of pitch on LPCC and PLPC and explore their efficacy for children's mismatched ASR in comparison to MFCC. The pitch of the signals is modified using the pitch synchronous time scaling (PSTS) method [10] reported to give faithful transformations for a wide range of transformation factors.

Section 2 describes the speech corpus and the experimental setup. Section 3 presents the study on the effect of pitch on LPCC and PLPC. Section 4 discusses children's mismatched ASR using different features with explicit pitch normalization. The paper is concluded in Section 5.

2. Speech Corpus and Experimental Setup

The connected digit recognizer used in this work is developed using HTK following the procedure explained in our previous work [5]. The training and test data for the digit recognizer are obtained from TIDIGITS database. The training set contains 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 contains 25,525 words from children of both sexes having age between 6-15 years and pitch values ranging from 100-360 Hz. All speech data is down sampled to 8 kHz. In this work, 'pitch' of speech signals refers to their average pitch estimated using the ESPS tool available in the Wavesurfer software package.

The 11 digits (0-9 and OH) are modeled as whole word left-to-right HMM using 16 states per word and 5 diagonal covariance Gaussian distributions per state. The speech is analyzed with a Hamming window of length 25 ms, frame rate of 100 Hz and pre-emphasis factor of 0.97. Both LPCC and PLPC features are derived using 12^{th} order LP analysis. A 21-channel triangular filterbank is used for PLPC computation. LPCC base features are 12-D (C_1 to C_{12}) while PLPC base features are 13-dimensional (13-D) (C_0 to C_{12}). In addition to the base features, their first and second order temporal derivatives, computed over a span of ± 2 frames, are also appended making the final feature dimension as 36 for LPCC and 39 for PLPC. Cepstral mean subtraction is also applied to all features. For ease of comparison, the results obtained in our previous study [5] for 39-D MFCC feature (C_0 to C_{12} base coefficients and their first and second derivatives) computed using 21-channel triangular filterbank are also given in this paper.

3. Effect of Pitch Variations on Features

In this section, we explore the effect of pitch variations across speech signals on LPCC and PLPC features in comparison to MFCC. Following the study done for MFCC in [5], LPCC and PLPC features are extracted for steady portions of 7 different vowels from the speech signals belonging to 'low' (100-125 Hz) and 'high' (200-250 Hz) pitch ranges selected from TIMIT database. Approximately, 2000 frames are used for

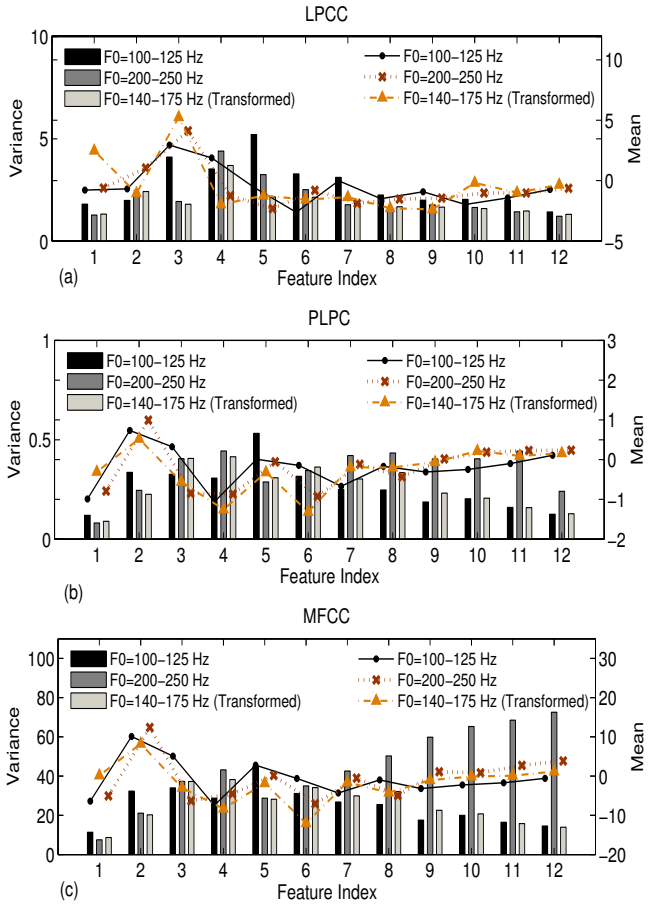

Figure 1: *Plots showing mean along with variance (in bar) of each of the coefficients (C_1 - C_{12}) of different features for signals of different pitch groups: 100-125 Hz, 200-250 Hz and 200-250 Hz transformed to 140-175 Hz for vowel /iy/ (a) LPCC (b) PLPC (c) MFCC.*

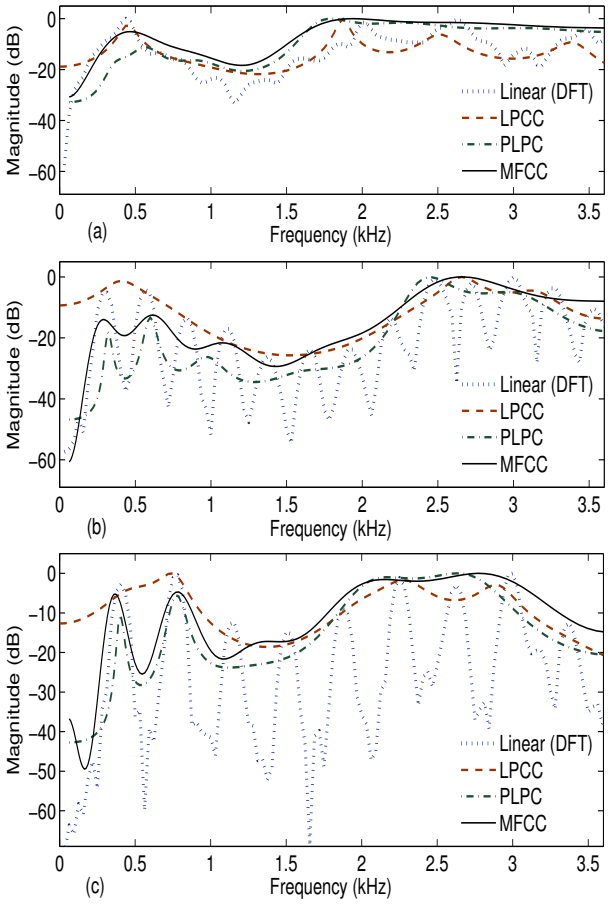

Figure 2: *Plots of smooth spectra corresponding to different features along with linear (DFT) spectrum for vowel /iy/ of pitch (a) 90 Hz (b) 220 Hz (c) 270 Hz.*

each vowel. The plots of mean and variance of each of the 12 base coefficients (C_1 - C_{12}) for a representative vowel /iy/ from 'low' and 'high' pitch group signals are shown in Fig. 1(a) and Fig. 1(b) for LPCC and PLPC, respectively. Similar plots obtained for MFCC in [5] are shown in Fig. 1(c). It is noted that, like in case of MFCC, the variances of the higher dimensions of PLPC of the 'high' pitch group signals are also significantly larger than those in case of the 'low' pitch group signals. However, the increase in the variances for PLPC is comparatively smaller than that in case of MFCC. On the other hand, in case of LPCC, no such significant change in the variances is observed. Further, the pitch of the 200-250 Hz pitch group signals is transformed to 140-175 Hz through a constant factor of 0.7 using PSTS method. The plots of mean and variance of each of the C_1 - C_{12} coefficients for vowel /iy/ from original and pitch transformed signals are also shown in Fig. 1(a) and Fig. 1(b) for LPCC and PLPC, respectively. It is noted that on pitch reduction the variances of the higher dimensions of PLPC also reduce considerably as noted in case of MFCC unlike those of LPCC.

In order to explore the effect of pitch on the smooth spectra corresponding to LPCC and PLPC in comparison to that observed for MFCC in [5], the smooth spectra corresponding to LPCC and PLPC are obtained for vowel /iy/ from signals with different pitch values by converting the cepstral coefficients to their corresponding LP coefficients and then obtaining the magnitude response of the all-pole filter for those LP coefficients and are shown in Fig. 2 along with their corresponding linear (DFT) spectrum and the smooth spectra corresponding to MFCC (from [5]). In case of PLPC, similar to MFCC, some pitch-dependent distortions are noted at lower frequencies (below 1.5 kHz) in the spectral envelope for 220 Hz and 270 Hz pitch signals when compared with that of the 90 Hz signal which are also found to increase with increasing pitch of the signal. Since, PLPC computation also incorporates a filterbank similar to that used for MFCC computation, we hypothesize that the possible cause for these pitch-dependent distortions in the spectral envelope for PLPC is also the insufficient smoothing of the pitch harmonics by the filterbank particularly by lower order filters of nearly 100 Hz bandwidth. This explains the cause of the earlier observed increase in the variances of PLPC with increasing pitch of the signals. In case of LPCC, slight pitch-dependent distortions in the spectral envelope are observed with pitch variations across the signals. These could be attributed to the aliasing distortions in the autocorrelation function for high pitch signals [8]. Since these distortions do not appear to be as severe as in case of MFCC and PLPC, this explains the earlier observed no significant change in the variances of LPCC for different pitch signals.

Table 1: *The mean and variance of MD of the original 'low' and 'high' and the pitch transformed 'high' pitch group signals for different features and vowels with respect to models trained on 75-100 Hz pitch group signals.*

Feature	Vowel	MD (Mean / Variance)		
		Pitch Group		
		Original		Transformed
		100-125 Hz ('low')	200-250 Hz ('high')	140-175 Hz (from 'low')
LPCC	/ae/	12.3 / 50.0	36.2 / 230.7	43.2 / 336.1
	/iy/	12.7 / 50.6	22.6 / 69.6	35.7 / 104.6
PLPC	/ae/	12.4 / 60.4	50.2 / 711.9	36.1 / 253.2
	/iy/	13.5 / 69.5	43.6 / 509.7	36.7 / 253.5
MFCC	/ae/	12.4 / 60.2	65.5 / 2553.1	35.3 / 282.6
	/iy/	13.5 / 85.3	59.2 / 1735.4	34.7 / 246.8

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

Feature	Condition	WER (%)			
		All F_o Values (7,772)	$F_o <$ 250 Hz (5,224)	$F_o =$ 250-300 Hz (2,346)	$F_o >$ 300 Hz (202)
LPCC	Baseline	20.29	15.85	25.79	46.72
	Norm.	20.06	15.78	25.28	46.59
PLPC	Baseline	10.61	6.70	15.52	33.09
	Norm.	10.05	6.44	14.62	30.61
MFCC	Baseline	11.37	6.54	17.47	39.03
	Norm.	9.64	6.02	14.24	30.11

Similar to the study done in [5], the mean and variance of Mahalanobis distance (MD) is computed for /ae/ and /iy/ vowels from 'low' and 'high' pitch group signals with respect to models trained on 75-100 Hz pitch group signals for both LPCC and PLPC and are given in Table 1 along with those for MFCC (from [5]). It is noted that though the mean and variance of MD increase with increasing pitch of the signals for both LPCC and PLPC but the increase is slightly smaller in case of PLPC while much smaller for LPCC in comparison to that for MFCC. On reducing the pitch of the 'high' pitch group signals by a factor of 0.7, a significant reduction in the mean and variance of MD is observed for PLPC for both vowels as noted in case of MFCC. However, for LPCC a small increase in the mean and variance of the MD is observed on pitch reduction which can be attributed to the small distortion introduced during the time-scaling operation in the pitch transformation algorithm. Although these distortions would also similarly effect the mean and variance of the MD in case of MFCC and PLPC, but with significant reduction in the pitch-dependent distortions with pitch modification net MD is reduced. Thus, the ASR performance should be though significantly effected by pitch variations in case of PLPC but little lesser than that for MFCC while it should not be significantly effected for LPCC.

4. Children's ASR Experiments

The recognition performances (in WER) on the adult test set for LPCC, PLPC and MFCC features are 1.02%, 0.47% and 0.43%, respectively. The baseline ASR performances on the children test set for the said three features are given in Table 2. Thus, adults' speech trained models give large degradation in the ASR performance for children's speech than for adults' speech for all the three features.

Children have much higher pitch values compared to those of adults. Motivated by the effect of pitch variations on the features observed in Section 3, the pitch of children's test speech is modified towards the pitch range of the adults' training data using PSTS method. For determining the appropriate pitch values for the pitch transformed signals, a 8-point ML grid search is done among the original signal and its seven pitch transformed versions with transformed pitch values ranging from 70-250 Hz in steps of 30 Hz as described in detail in [5]. The recognition performance of the pitch normalized children test set for the said three features are also given in Table 2.

From Table 2, it is noted that with explicit pitch normalization PLPC gives relative improvement of 5% while MFCC gives 15% relative improvement over their corresponding baselines for children's mismatched ASR which are attributed to reduction of the pitch-dependent distortions in the spectral envelope. The lesser improvement in the ASR performance with pitch normalization using PLPC than that obtained with MFCC is consistent with our earlier hypothesis and observations. However, it is to note that though PLPC gives better baseline performance than MFCC for children's ASR under mismatched condition due to equal-loudness pre-emphasis and cubic-root amplitude compression in PLPC computation [11], after explicit pitch normalization of children's speech MFCC results in better ASR performance than PLPC. Observing the pitch-group wise performances given in Table 2, it is noted that with pitch normalization more consistent improvements are obtained for different pitch groups i.e., higher pitch groups show greater improvements in case of MFCC than for PLPC. As PLPC incorporates equal-loudness pre-emphasis [11] which deemphasizes the 0.4-1.2 kHz spectral region by 6dB compared to the 1.2-3.1 kHz spectral region, the effect of pitch-dependent variations (appearing mainly below 1.5 kHz) is significantly reduced. However, this may also lead to de-emphasis of the relevant spectral information which becomes more obvious when the test data is explicitly pitch normalized or devoid of significant pitch differences. This explains the less consistent improvements obtained for different pitch groups and slightly impaired children's ASR performance after explicit pitch normalization in case of PLPC. The above fact is also supported by earlier observation that with no significant pitch differences between training and test sets, the performance of the adult test set is found to be slightly inferior for PLPC compared to that for MFCC. Therefore, MFCC is a better feature for children's ASR under mismatched condition than PLPC.

In case of LPCC, an insignificant improvement in the ASR performance is observed after pitch normalization which is consistent with our earlier hypothesis and observations indicating greater pitch-robustness of LPCC than MFCC. The slight improvement in the children's mismatched ASR performance with pitch normalization could be attributed to the reduction of the aliasing distortions in the LP spectral envelope due to reduction of pitch [8]. However, the performance obtained by LPCC is significantly poor than that obtained with MFCC even after pitch normalization as observed in case of their corresponding baseline performances. Thus, despite being more pitch robust, LPCC is a poor feature for children's mismatched ASR than MFCC.

5. Conclusions

In this work, the effect of pitch variations across speech signals on LPCC and PLPC features has been studied to explore their efficacy for children's speech recognition on adults' speech trained models in comparison to MFCC feature. The study shows that, though the variances of coefficients of PLPC and their corresponding smoothed spectral envelope are effected similar to those of MFCC in case of high pitch signals due to insufficient smoothing of the pitch harmonics in the speech spectrum by the filterbank, but comparatively to a lesser extent. On the other hand, unlike MFCC, no such major effects of pitch variations are noted in case of LPCC. However, on observing the children's ASR performances under mismatched condition using LPCC, PLPC and MFCC features, though PLPC is noted to give the best baseline performance but after explicit pitch normalization of children's speech, the best ASR performance for children's speech on adults' speech trained models is obtained with MFCC.

6. References

[1] Singer, H. and Sagayama, S.,"Pitch dependent phone modelling for HMM based speech recognition", in Proc. ICASSP, 1:273–276, March 1992.

[2] Xu Shao and Milner, B.,"Pitch prediction from MFCC vectors for speech reconstruction", in Proc. ICASSP, 1:I–97–I–100, May 2004.

[3] Gustafson, J. and Sjölander, K.,"Voice transformations for improving children's speech recognition in a publicly available dialogue system", in Proc. ICSLP, 297–300, September 2002.

[4] Dharanipragada, S and Yapanel, U. H. and Rao, B. D.,"Robust feature extraction for continuous speech recognition using the MVDR spectrum estimation method", IEEE Trans. on Audio, Speech, And Language Processing, 15(1):224–234, January 2007.

[5] Sinha, R. and Ghai, S.,"On the use of pitch normalization for improving children's speech recognition", in Proc. Interspeech, 568–571, September 2009.

[6] Ghai, S. and Sinha, R.,"Analyzing pitch robustness of PMVDR and MFCC features for children's speech recognition", in Proc. SPCOM, 1–5, July 2010.

[7] Yapanel, U. H. and Hansen, J. H. L.,"A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition", Speech Communication, 50(2):142–152, February 2008.

[8] Rahman, M. S. and Shimamura, T.,"Linear prediction using refined autocorrelation function", EURASIP Journal on Audio, Speech and Music Processing, 2007:1–9, Article ID 45962, 2007, doi:10.1155/2007/45962.

[9] Psutka, J. and Müller, L. and Psutka, J. V.,"Comparison of MFCC and PLP parameterizations in the speaker independent continuous speech recognition task", in Proc. Eurospeech, 1813–1816, 2001.

[10] Cabral, J. P. and Oliveira, L. C.,"Pitch-synchronous time-scaling for prosodic and voice quality transformations", in Proc. Interspeech, 1137–1140, 2005.

[11] Hermansky, H.,"Perceptual linear predictive (PLP) analysis of speech", J. Acoust. Soc. Amer., 87(4):1738–1752, April 1990.