Pitch Mean Based Frequency Warping

Jian Liu, Thomas Fang Zheng, and Wenhu Wu

Center for Speech Technology, Tsinghua National laboratory for Information Science and Technology, Tsinghua University, Beijing, 100084
liuj@cst.cs.tsinghua.edu.cn, {fzheng, wuwh}@tsinghua.edu.cn

Abstract. In this paper, a novel pitch mean based frequency warping (PMFW) method is proposed to reduce the pitch variability in speech signals at the front-end of speech recognition. The warp factors used in this process are calculated based on the average pitch of a speech segment. Two functions to describe the relations between the frequency warping factor and the pitch mean are defined and compared. We use a simple method to perform frequency warping in the Mel-filter bank frequencies based on different warping factors. To solve the problem of mismatch in bandwidth between the original and the warped spectra, the Mel-filters selection strategy is proposed. At last, the PMFW mel-frequency cepstral coefficient (MFCC) is extracted based on the regular MFCC with several modifications. Experimental results show that the new PMFW MFCCs are more distinctive than the regular MFCCs.

Keywords: Pitch, frequency warping, MFCC.

1 Introduction

State-of-the-art speech recognition systems have to face with a lot of variability in the acoustic signal. For example, context variability, style variability, speaker variability and environment variability are some typical types of variability [1] that may cause mismatch between training and test data of automatic speech recognition (ASR) systems. A lot of schemes have been developed in the past few years to compensate for this mismatch in order to improve the accuracy of the ASR system. Two major schemes are acoustic features transformation and acoustic model parameters adaptation.

Speaker variability and speaking style variability are two major factors that might cause mismatch between the trained acoustic models and the actual speech to be recognized. To reduce speaker variability, vocal tract length normalization [2, 3, 4] is commonly used to transform acoustic feature for ASR. The correlation between a speaker's average pitch and the vocal tract length was also exploited in [5]. On the other hand, even for a same speaker, his/her speech will change much with different speaking styles, therefore the speaking style variability also need to be considered. The pitch contour is one of the important features to classify different speaking style. There is a correlation between the pitch contour and the speaking style. Thus both the speaker variability and the speaking style variability correlate with the pitch frequency.

Q. Huo et al. (Eds.): ISCSLP 2006, LNAI 4274, pp. 87 – 94, 2006.

Automatic speech recognition should be based on speech features that contain relevant information capable of discriminating different speech sounds. The dynamic features of pitch are useful in speech recognition, especially for Chinese. Yet speech signals with different average pitches could contain the same phonetic information. Even the phase vocoder can manipulate the signal in frequency-domain, enabling pitch-shifting without changing the phonetic information [6, 7]. Thus the average pitch of speech is not the relevant feature to discriminate different phonetic information in speech for ASR. Commonly used speech features such as MFCCs are affected by changes of the pitch in speech signals. One way to alleviate the disturbance of pitch is to find speech features that are less sensitive to changes of pitch, yet capable of retaining good discriminative properties.

In this paper, a pitch mean based frequency warping (PMFW) method is proposed in feature extraction to compensate for the pitch-mismatch in speech signals. In [8, 9], the formant-based frequency warping was discussed for speaker normalization. However, the motivation of this paper is not only implementing speaker normalization. Because the average pitch is not directly proportional to the vocal tract length [5], using pitch explicitly for speaker normalization is not so reasonable as expected. On the other hand, the average pitch of the speech segment does have relations to both the speakers and the speaking styles. Effects of the pitch-mismatch need to be considered separately in ASR. Our work presents an approach that warps the frequency according to the average pitch of a speech segment. The motivation here is to integrate the PMFW into the acoustic feature extraction at the front-end with a little computation and make the new PMFW features more discriminative for ASR.

2 Pitch Mean Based Frequency Warping

2.1 ASR and Pitch

Pitch plays an important role in speech perception. The pitch is not a characteristic of the vocal tract length and does not directly affect the resonant frequencies. However, the information about pitch can be used to improve ASR systems. There are three typical methods that use the pitch information in ASR systems.

First, the pitch can be used as an acoustic feature and modeled using hidden Markov models (HMMs) and/or Artificial Neural Network (ANN). For example, in [10], the dependency between the hidden state and the pitch was modeled implicitly. The ASR system could achieve significant improvement by incorporating the pitch frequency.

Second, the pitch can be used to synchronize the frame size and/or the shift. A constant frame size and a constant shift are always used in ASR systems. The power spectral estimation may include artifacts without aligning the frames to the natural pitch cycles. A pseudo pitch synchronous method was proposed in [11] which improved the robustness and accuracy for low SNR speech.

Third, the pitch can be used for frequency warping factor estimation. In [5, 12], the correlation between a speaker's average pitch and the vocal tract length was exploited and the probability distribution of warp factors conditioned on pitch observations was modeled. That pitch-based warp factor estimation can be an effective method of improving ASR performance.

In ASR systems, the MFCC is one of major acoustic features. MFCC features are calculated from the power spectrum, and include some harmonic structure related to the pitch. Variations in pitch could cause variations in features. As a result the pitch mismatch and the variability in features have effects on speech recognition systems. On the other hand, the pitch variability and the speaker variability do not have direct relations. Even the same speaker could have pitch variability, such as at different mental conditions. The variability due to pitch will be implicitly alleviated by training a speech recognition system on a corpus collected from a large, diverse collection of speakers. However, the explicit reduction or elimination of pitch-included feature variability could lead to better recognition performance.

2.2 Pitch Mean Based Frequency Warping

Frequency warping is a typical kind of methods for feature transformation. The frequency axis is scaled by a warping function $f_{\alpha}(\omega)$, where α is a warping factor. Given the power spectrum, $X(\omega)$, of a speech signal, the warped spectrum is

$$Y(\omega) = X\left(f_{\alpha}\left(\omega\right)\right) \tag{1}$$

The warping function $f_{\alpha}(\omega)$ is always assumed invertible, i.e. strictly monotonic and continuous [3]. The warping function should conserve the bandwidth and information contained in the original spectrum in theory. However, there is redundant information in the original spectrum and only a subband of spectrum is useful for frequency warping. In our work, a linear frequency function is used, i.e. $f_{\alpha}(\omega) = \alpha\ \omega$. The reason for using a linear frequency function is that it has explicit physical meaning. According to the Fourier transformation, the compressing or stretching in frequency axis is equivalent to the re-sampling of the waveform in time axis, i.e. $X(\alpha\ \omega) = \frac{1}{\alpha} x(t/\alpha)$. Thus warping frequency with a linear function could alleviate the pitch-mismatch in the speech signal. Generally speaking, the phonetic information is '*hidden*' in the relative spectrum. The frequency warping adjusts the spectrum to determine more distinctive bands for ASR in some sense. It has been proved that the Maximum-Likelihood (ML) based frequency warping is effective for ASR [2], however, it requires more data and computation. Because perceiving pitch is natural for human and human can process speech properly with pitch variations, such as singing and speaking, we will focus on the relations between the pitch and the warping factor.

How to determine the warping factor α is important for frequency warping. In our method, the warping factor is dependent of the average pitch of a certain speech segment. We can assume that the warping factor α is a function of the average pitch as follows

$$\alpha = g\left(p\right) \tag{2}$$

where p is the pitch mean of a speech segment. Our goal here is to determine an analytic approach to expressing the relationship between the warping factor and the average pitch. Thus, two monotonic and continuous functions are exploited and compared in our experiments. Two typical functions are defined as follows

$$g(p) = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min}) \bullet \frac{(p - p_{\min})}{(p_{\max} - p_{\min})} \tag{3}$$

$$g(p) = 1 + \frac{(\alpha_{\max} - \alpha_{\min})}{2} \bullet \frac{\log_2\left(p^2/(p_{\min} \bullet p_{\max})\right)}{\log_2\left(p_{\max}/p_{\min}\right)} \tag{4}$$

where $p_{\min}$ and $p_{\max}$ are the minimal and maximal pitch values of the human voice, respectively, $\alpha_{\min}$ and $\alpha_{\max}$ are the lowest and highest bounds of the warping factor, and p is the pitch mean of a speech segment. Empirically the pitch range of human voice is from 50 Hz to 500 Hz approximately. In our experiments, the range of the pitch is from 55 Hz to 440 Hz for convenience and any pitch with its value lower (or higher) than 55 (or 440) Hz is set to 55 (or 440) Hz. $\alpha_{\min}$ (or $\alpha_{\max}$) is also determined empirically as 0.85 (or 1.15) in our experiments. Equation (3) is in a linear form meaning that the warping factor is proportional to the average pitch in a linear space while Equation (4) is in a nonlinear form meaning that the warping factor is proportional to the average pitch in an octave space.

The pitch mean of the speech segment is calculated as

$$p = \frac{1}{N} \sum_{\substack{t=0 \\ 55 \le p_t \le 440}}^{t=T} p_t \tag{5}$$

where T is the total frame number of a speech segment, p_t is the pitch value at frame t (if no pitch value is successfully estimated at frame t, p_t is set to 0 in the above equation), and N is the total number of frames at which pitch value is successfully estimated. Note that T could be set as a fixed time period such as 2 seconds, however, for convenience, in our experiments, each sentence is considered as a speech segment and T is set to its length.

In practice, errors in pitch estimation are inevitable. However the average of pitch in a speech segment can be calculated with little bias if the speech segment is long enough. The method in [13] can be used to balance the doubling error rates and the halving error rates in pitch estimation to get more accuracy pitch mean in a speech segment.

2.3 PMFW Derived Feature

The proposed PMFW is integrated in the feature extraction at the front-end without additional computation in the training and the decoding procedures. The PMFW features are based on the standard MFCCs with two additional steps added as shown in Fig. 1.

First, the pitch mean based frequency warping is performed in the Mel-filters frequencies. The frequency warping can be implemented by simply varying the spacing and width of the component filters of the filter bank without changing the original speech signal [2]. The PMFW is implemented as follows

$$B'(k) = \alpha B(k), \quad k = 0,1,...,N+1 \tag{6}$$

where $B(k)$ is the start frequency of Mel filter k (for example, $B(0)$, $B(1)$, $B(2)$ are start, middle and end frequencies of the first Mel filter, respectively), and N is the total number of Mel filters. Equation (6) means that a male speaker with a smaller α would use a relative low band of frequency to calculate features, and vice versa. It can be assumed that male speakers' phonetic information is hidden in relative low frequency bands and female speakers' is hidden in relative high frequency bands.

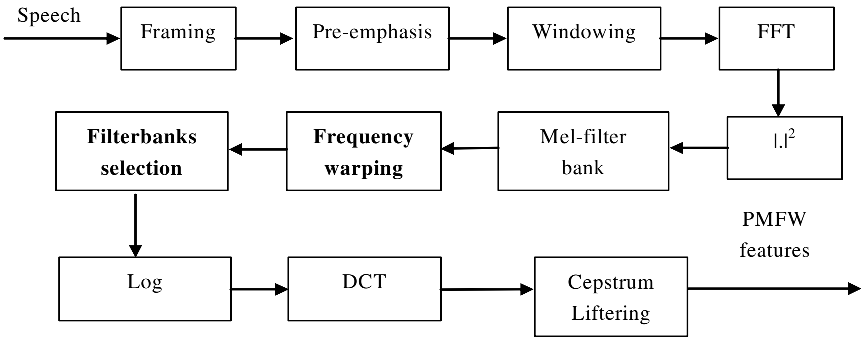

Fig. 1. Schematic diagram for PMFW feature extraction

Second, Mel filters are selected. Different warping factors will bring mismatch in bandwidth between different speech signals. In [2, 3], the piecewise warping functions were used to solve this problem. Using the piecewise warping function ensures that the full frequency band is used in features at different warping factors. However, the full band is not used in our method. We determine stable sub bands by cutting off a fixed number of lowest/highest bands at different warping factors. The selected sub bands should contain the same number of Mel filters. The number of filters that should be cut off is determined by

$$n = \underset{\alpha_{\max} B(k) \le f_{\max}}{\arg\max} \ \alpha_{\max} B(k) \tag{7}$$

where $f_{\max}$ denotes the maximal signal bandwidth. The filter start frequencies selected after PMFW are $B'(N+1\text{-}n)$, …, $B'(n)$. Both the lower and the higher filters will be cut off. For the experiments described in this paper, the sampling rate is fixed at 16 kHz, imposing a limit on the maximum signal bandwidth of 8 kHz. 35 (N=35) Mel filters are used and n=34 calculated by using Equation (7). Thus $B'(2)$, …, $B'(34)$ are selected to use for feature extraction after PMFW.

3 Experiments and Discussion

3.1 Experimental Setup

Experiments were designed to compare the performance of systems using traditional MFCCs and PMFW MFCCs. A subset of 863CSL corpus [14], which is a continuous speech database, was used in our experiments. The training set contained 20 male speakers' data and 20 female speakers' data, totally 21,749 sentences (10,824 sentences for male and 10,925 for females) for about 22 hours. The test set contained another 8 male speakers' data and 8 female speakers' data, totally 8,941 sentences (4,524 for males and 4,417 for females) for about 9 hours.

In our experiments, we used HTK version 3.2 [15] for training, testing, and the baseline's MFCCs feature extraction. The PMFW MFCCs were extracted by our own program using the algorithm proposed in this paper. The pitch, MFCCS and PMFW MFCCs were extracted every 12 milliseconds. Both PMFW MFCCs and traditional MFCCs were 26 dimensional, consisting of 13 static coefficients and corresponding

Table 1. Recognition results with gender matched/mismatched training and test data ('Linear' means using Equation (3) to calculate the warping factor while 'Octave' using Equation (4) to calculate the warping factor; *n* M (or *n* F) means the training or testing set contains speech data by *n* male (or female) speakers; the performance is evaluated in syllable accuracy rate, in %)

Training Set / Test Set	Method	20 M (%)	20 F (%)
8 M	Baseline	67.69	15.42
	Linear	69.90	48.59
	Octave	69.31	44.60
8 F	Baseline	20.71	78.90
	Linear	53.87	81.03
	Octave	48.78	80.56

Table 2. Recognition results with gender independent training (*n* M (or *n* F) means the training or testing set contains speech data by *n* male (or female) speakers; the performance is evaluated in syllable accuracy rate, in %))

Training Set / Test Set	Method	20 M +20 F (%)
8 M	Baseline	67.27
	Linear	70.99
	Octave	70.02
8 F	Baseline	78.04
	Linear	80.77
	Octave	80.19
8 M + 8 F	Baseline	72.59
	Linear	75.82
	Octave	75.04

13 delta coefficients. The 5-state 8-mixture Hidden Markov Model (HMM) topology was adopted to model the toneless tri-IFs where IF means either a Chinese initial or a Chinese final. The speech recognition units were 397 Chinese syllables (with tone disregarded).

3.2 Results and Discussion

The first experiment was designed to compare the performance between traditional MFCCs and the PMFW MFCCs on gender dependent models. Table 1 illustrates that when the traditional MFCCs are used, there will be very large drops in syllable accuracy rate when there is gender mismatch between the training speakers and the testing speakers. When the PMFW MFCCs are used, the accuracy rate will be improved considerably no matter the training speakers and the test speakers match in gender or not.

Traditional MFCCs perform badly when the training and test speakers gender mismatch. Thus some ASR systems use gender dependent models and perform gender recognition before speech recognition. The PMFW MFCCs could alleviate variations caused by gender mismatch. Although the accuracy rates in the gender mismatch test are lower than that in the gender matched test when PMFW MFCCs used, the accuracy will be remarkably increased in contrast to traditional MFCCs. Comparing two functions for calculating the warping factor, we can see that using a linear function to restrict the warping factor and the pitch mean could achieve higher accuracy at all tests in contrast to the octave function.

The second experiment was designed to compare the performance between traditional MFCCs and the PMFW MFCCs on the gender independent models. The percentages correct for the baseline when tested with 8 male and 8 female speakers were 67.27% and 78.04%, respectively, which were lower than those (67.69% and 78.90%, respectively) in the first experiment. The size of the training set in the second experiment was about twice larger than that in the first experiment, however, the accuracy rates were lower. The reason for that could be that traditional MFCCs of male and female speech are relatively diverged in the feature space, although the acoustic models used in two experiments both were 5-state 8-mixture based HMMs. Thus, it might be more difficult to model the distributions of the gender independent features than the gender dependent features when the acoustic model parameter size is fixed.

According to the first experiment, the linear function for warping factor calculation had better performance, so here we will only discuss the results when using the linear function here. When using PMFW MFCCs, the percentages correct when tested with 8 male and 8 female speakers were 70.99% and 80.77%, respectively, in the second experiment, and 69.90% and 81.03%, respectively, in the first experiment. It shows that the accuracy for males has been increased by 1.09%, while that for females has been decreased by 0.26%. Compared with traditional MFCCs, the PMFW MFCCs could have better performance, in other words, the PMFW MFCCs can have more convergence in the feature space than traditional MFCCs. Furthermore, in 16-speaker test (8 male and 8 female), there was a syllable error rate reduction of 11.8% when linear PMFW MFCCs were used in contrast to the traditional MFCCs.

4 Conclusion

The motivation of this paper is to extract more distinctive features at the front-end with little extra computation. By exploiting the correlation between the pitch and the speech, we propose an effective pitch mean based frequency warping method. To alleviate the pitch variations in speech signals, the warping factor is considered as a function of the average pitch of a speech segment. Then, two typical functions of the pitch mean are defined to calculate the warping factor. Furthermore, a simple method for performing frequency warping in the Mel-filter bank frequencies is described. The Mel filters selection strategy is presented for solving the mismatch in bandwidth between the original and the warped spectrum. Based on these operations, the PMFW MFCCs is extracted instead of the traditional MFCCs. Experimental results show that the PMFW MFCCs have better performance than traditional MFCCs

References

1. Huang, X. D., Acero A., Hon H. W.. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall, New Jersey, 2001.
2. Lee L., Rose R. C.. "Speaker Normalization Using Effiecient Frequency Warping Procedures," *Proc. ICASSP*, 353-356, 1996.
3. Pitz M., Ney H.. "Vocal Tract Normalization as Linear Transformation of MFCC," *Proc. EUROSPEECH*, 1445-1448, 2003.
4. Wang W., Zahorian S. A.. "Vocal Tract Normalization Based on Spectral Warping," *Proc. ICSLP*, 1185-1188, 2004.
5. Faria A., Gelbart D.. "Efficient Pitch-based Estimation of VTLN Warp Factors," *Proc. INTERSPEECH*, 213-216, 2005.
6. Laroche J., Dolson M.. " New Phase-Vocoder Techniques for Pitch-Shifting, Harmonizing and Other Exotic Effects," *Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustic*, 91-94, 1999.
7. http://www.dspdimension.com/data/index.html
8. Gouva E. B., Stern R. M.. "Speaker Normalization through Formant-based Warping of The Frequency Scale," *Proc. EUROSPEECH*, 1139-1142, 1997.
9. Zhan P., Westphal M.. "Speaker Normalization Based on Frequency Warping," *Proc. ICASSP*, 1039-1042, 1997.
10. Magimai-Doss M., Stephenson T. A., Bourlard H.. "Using Pitch Frequency Information in Speech Recognition," *Proc. EUROSPEECH*, 2525-2528, 2003.
11. Zilca R. D., Kingsbury B., et al. "Pseudo Pitch Synchronous Analysis of Speech with Applications to Speaker Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):467-478, 2006.
12. Glavitsch U.. "Speaker normalization with respect to F_0: a perceptual approach," *TIK-Report Nr. 185*, 2005.
13. Liu J., Zheng T. F., et al. "Real-time Pitch Tracking Based on Combined SMDSF," *Proc. INTERSPEECH*, 301-304, 2005
14. Wang D., Zhu X.Y., Liu Y.. "Multi-Layer Channel Normalization for Frequency-Dynamic Feature Extraction," Journal of Software, v 12, n 9, September, 2005, p1523-1529
15. Young S., et al. *The HTK book (for HTK version 3.2).* Cambridge University Engineering Department, 2002