International Conference on Signal Pocessing, pp. II-756~759, Aug. 21-25, 2000, Beijing

REFERENCE POINT ALIGNMENT FREQUENCY WARP METHOD FOR SPEAKER ADAPTATION

Tranzai LEE, Fang ZHENG, Wenhu WU
Speech Laboratory, Department of Computer Science and Technology,
Tsinghua University, Beijing 100084, P.R. of China
tranzai@sp.cs.tsinghua.edu.cn

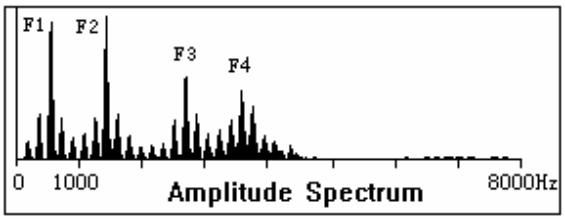
ABSTRACT

The variations of speakers' vocal tract shapes result in the variations of the formant positions and sequentially in the variances of the features extracted from every frame of speech. In order to remove or reduce the variations of the formant positions, a speaker adaptation method will be proposed and investigated in this paper which is based on a frequency warp function (*fwf*). The *fwf* warps the frequency axis so that the variations can be reduced. For a given speaker, some frequency reference points are selected to help to get this *fwf* by finding the relationship between the positions of these reference points before and after the warping. According to the new positions of those reference points for the given speaker, the *fwf* can then be constructed. The experimental results show that this method reduces the error rate by an average of 14.5%.

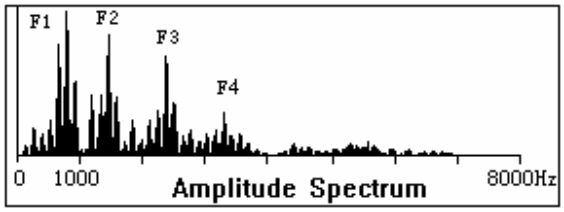
1. INTRODUCTION

Though great progress has been made in continuous speech recognition (CSR) in recent years, a problem is still always being encountered, i.e., the accuracy of the speaker independent (SI) systems will reduce greatly in comparison with the speaker dependent (SD) ones. This problem comes from the variations among speakers and sequentially from the variations of the features extracted from every frame of speech among speakers. To resolve this problem, some adaptation techniques are proposed and adopted, which are proved to have taken effect. Some of them, such as the vocal tract length (VTL) normalization methods ([1], [2], [3], [4], [5]), can be used in the stage of feature extraction. A typical method is the linear frequency warp method proposed for speaker normalization [1]. This method is based on the standpoint that the speech spectrum of one speaker is stretched or compressed linearly from that of another speaker and hence the two speaker's difference lies in the VTL only. In that case, the VTL normalization can be performed by linearly warping the frequency axis for a speaker's speech spectrum. Some other adaptation techniques can be used after features are extracted. A typical method is the maximum likelihood linear regression (MLLR) method [6] designed for the speaker adaptation of mixture Gaussian based hidden Markov models (HMM). This method is based on the fact that speakers' variations lead to the result that features of different speakers occupy different positions in the feature space. Therefore, in order to remove the variations among different speakers, linear transformations are performed to shift every speaker's feature space to the same position for the Gaussian mixture based hidden Markov models.

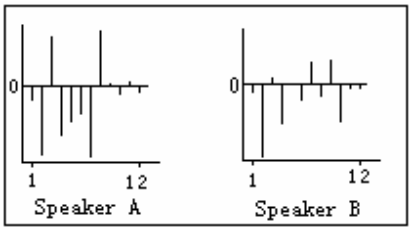
In the next section, we will illustrate the variations existing among the speakers. We are inspired by this example and give a speaker adaptation method in section 3. In section 4, some experiments are designed to test the effect of this method.



(a) Spectrum for Speaker A



(b) Spectrum for Speaker B



(c) 12-order mel-cepstrum for Speaker A and Speaker B

Fig.1 (a), (b) the amplitude spectrums for the Mandarin vowel [e] calculated from the data uttered by Speakers A and B respectively. (c) The mel-cepstrums obtained from the amplitude spectrums in (a) and (b).

2. VARIATIONS AMONG DIFFERENT SPEAKERS

Among speakers, the variations come mainly from the variations of vocal tract shapes and vocal cords. Those variations will lead

to the variations of the speech formants' positions. Even though for the same sonant, the positions of the corresponding formants still vary among different speakers. For a uniform tube, the formants' positions only depend on the tube's length. The VTL varies among different speakers. But it is possible that the formant positions of different speakers vary in the way that is inconsistent to their VTLs' variances when uttering the same sonant (Fig.1).

According to two speaker's amplitude spectrum showed in Fig.1, we can be obvious that the fundamental frequency of Speaker A is higher than that of Speaker B. The frequency of Speaker A's first formant, F_1, is higher than that of Speaker B's first formant, F_1. If only taking the position of the first formant into consideration, we can infer that the VTL of Speaker A is longer than that of Speaker B. However, the frequencies of Speaker A's the third formant, F_3, and the fourth formant, F_4, are higher than the corresponding frequencies of Speaker B's formants F_3 and F_4. Then, we will come to an opposite conclusion that the VTL of Speaker A is shorter than that of Speaker B if taking the formants F_3 and F_4 into consideration. According to the above discussion, it is easy to understand why their mel-cepstrums (Fig.1(c)) are different.

From the above example, we can come to the conclusion that the positions of the corresponding formants among speakers vary in the way that is not always consistent to their VTLs' variations for the same sonant. As a result, it is possible that a speaker's formant positions can not be linearly mapped to another speaker's positions. This phenomenon can be explained as follows. The different vocal tract shapes produce different responses to the same frequency. Therefore, different vocal tract shapes lead to different frequency warps, and different frequency warps result in different formant positions. The formant positions are determined according to the vocal tract shape including the VTL.

The linear frequency warp method mentioned in [1] only linearly stretches or compresses the speech spectrum. Therefore, according to above discussion, this speaker normalization method will not work well if applied to previous example. In another hand, the transformation from the frequency spectrum to the mel-cepstrum is nonlinear. The corresponding formant positions' variations between two speakers may be nonlinear, so the variations between the mel-cepstrums of two speakers can not be removed by a linear transformation from one speaker's mel-cepstrum space to another speaker's mel-cepstrum space. In our opinion, the MLLR method [6] is not so perfect for the speaker adaptation, at least for the above example.

For the sake of the speaker adaptation, it is significant to align the corresponding formants of the same sonant of different speakers. We can do it by the frequency warping. At first, we find the characteristics of the frequency warp for a given speaker and get a frequency warp function (abbreviated as *fwf*):

$$f' = warp(f) \tag{2.1}$$

Where f is the speaker's frequency while f' the rectified frequency. The mel-cepstrum is extracted in the domain of f'. The function $warp(f)$ must be monotone, one-to-one function within the interval [0, f_{max}], where f_{max} is the maximum frequency (i.e., the cut-off frequency) of the given signal while $2f_{max}$ the sampling rate. Function $warp(f)$ maps the interval [0, f_{max}] into the interval [0, f_{max}] and there are at least two fixed points, 0 and f_{max}, in this mapping. Then, the speaker's formant positions are rectified by means of the speaker's *fwf*. Unfortunately, it is not easy to find all the corresponding *fwfs* for every different sonant of a given speaker. However, we can get a simplified version under the assumption that the frequency warp fashions are approximately consistent for different sonant units of the same speaker. Hence a unique *fwf* can be used for a single speaker. As a matter of fact, it is still very difficult to find a unique and consistent *fwf* for a given speaker. In this paper, we propose a reference point alignment frequency warp method, an easy way to find the *fwf*, to facilitate the speaker adaptation. In this method, we align the corresponding formants for different speaker by aligning a set of reference points.

3. REFERENCE POINT ALIGNMENT FREQUENCY WARP METHOD

In this section, we give the method to find the *fwf* specific to a given SI speech recognition system for a given speaker.

Given an SI system, we suppose there is a standard speaker, say Speaker A, for whom the SI system has a highest recognition accuracy. N-1 points, f_1 through f_{N-1}, are selected in the interval (0, f_{max}) as the reference points. For any other given speaker, say Speaker B, we assume that the corresponding formants of Speaker A and Speaker B do not match well owing to speaker B's frequency warp. As a result, the reference points, f_1 through f_{N-1}, are shifted to new positions, $f_1^b, f_2^b, \cdots, f_{N-1}^b$, respectively, for speaker B. According to our previous assumption, the distributions of points, $f_1^b, f_2^b, \cdots, f_{N-1}^b$, are approximately consistent for different sonant units of the same speaker, that is to say, every single speaker has its unique reference frequency point distribution. Once the points $f_1^b, f_2^b, \cdots, f_{N-1}^b$ are obtained, we can get the reference point alignment frequency warp function related to Speaker B as

$$warp(f) = f_i + \frac{f_{i+1} - f_i}{f_{i+1}^b - f_i^b}(f - f_i^b), if \ \ f \in (f_i^b, f_{i+1}^b) \quad 0 \le i \le N \tag{3.1}$$

Where

$$f_0^b = f_0 = 0, \tag{3.2}$$

and

$$f_N^b = f_N = f_{\max} \tag{3.3}$$

By means of this frequency warp function, the corresponding formants of Speaker A and Speaker B will be aligned. We refer

to this method as the *Reference Points Alignment* (RPA) frequency warp method.

The frequency warp function *warp(f)* is piecewise linear because the warping is linear between any two adjoining reference points. If we take N=2 (only one reference point, f_1), this method looks like the piecewise linear frequency warp method mentioned in [1]. But they have essential difference. The piecewise linear frequency warp method requires that the point f_1 should fall above the highest significant formant. For the example showed in Fig.1, it means that f_1 must be higher than the formant F_4. Then, the corresponding formants of two speakers will not be aligned using the piecewise linear frequency warp method mentioned in [1] for the example showed in Fig.1.

For the given Speaker B, a difficult problem remained is how to find these reference points $f_1^b, f_2^b, \cdots, f_{N-1}^b$. Maybe the analytical way is better, but it is not available up to now. We can perform the frequency warp search by the numerical fashion instead. This search process is explained as follows.

Step 0:

X : Speaker B's utterance; N-1: the number of reference points; m: the number of search step for finding a reference point.

Step 1:

Let i=N-1 and assume that $f_{i+1}^b, \cdots, f_{N-1}^b$ have been obtained.

Step 2: finding f_i^b .

Let $\Delta = \dfrac{f_{i-1} - f_{i+1}^b}{m}$.

FOR k=1,..., m-1 DO

$$f_{i,k}^b = f_{i+1}^b - k\Delta$$

$$f_{j,k}^b = \frac{f_{i,k}^b}{f_b} f_j, \quad j = 1, \cdots, i-1$$

Taking $f_{i,k}^b, \cdots, f_{i-1,k}^b, f_{i,k}^b, f_{i+1}^b, \cdots, f_{N-1}^b$ as Speaker B's reference points, which forms an instance of the *fwf* for Speaker B, as described in Equation (3.1). And sequentially, the evaluation score $s_k(X)$ of this *fwf* in the k-th iteration could be calculated, which will be described later.

ENDFOR

Then,

$$k^* = \arg\max_k s_k(X)$$

and

$$f_i^b = f_{i+1}^b - k^*\Delta .$$

Step 3 :

if (i>1) then i=i+1; go to step 2.

else search end.

The evaluation score $s_k(X)$ can be chosen as the recognition accuracy, or the probability $p(w_{i_1} \cdots w_{i_M} | X)$ for the hybrid HMM/ANN based system ([7], [8]), or the probability $p(X | w_{i_1} \cdots w_{i_M})$ for HMM based system, where $w_{i_1} \cdots w_{i_M}$ is the X's word string containing M words.

Because not all distributions of reference points are searched, it is not guaranteed that the optimal reference points can be found in this search process. Only sub-optimal results are definitely attainable.

4. EXPERIMENTS

To test our proposed reference points alignment frequency warp method, some experiments are designed and done. Our baseline system is the speaker independent Mandarin continuous speech recognition system based on the hybrid HMM/ANN method. In this system, the language model is not used and hence only acoustic recognition results are given. For every speaker, There are 520 ~ 650 Mandarin utterances. 50 utterances of them are used for training and others for testing. In our first experiment, f_{max} is 8kHz and 8 reference points, i.e., 1kHz, 2kHz, 3kHz, 4kHz, 5kHz, 6kHz, 7kHz, and 7.9kHz, are selected. The experimental result is showed in Tab.1.

Tab. 1. The comparison of the error rates without and with the RPA frequency warp. 8 reference points are used.

No. of speaker	Error rate (Baseline)	Error rate (with RPA)	Error Reduce
1	28.2%	25.1%	11.0%
2	52.0%	38.9%	25.1%
3	47.3%	34.2%	27.6%
4	23.5%	20.7%	11.9%
5	29.6%	26.2%	11.5%
6	27.8%	26.3%	5.3%
7	49.7%	45.3%	8.8%
Average	---	---	14.45%

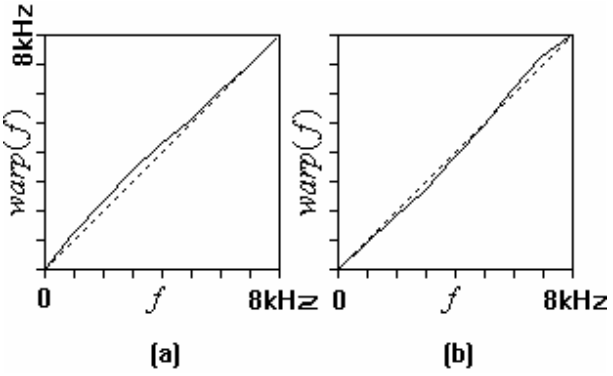


Fig.2. The frequency warp functions (*fwf*) obtained by using 8 reference points. (a) Speaker 1 in Tab.1; (b) Speaker 2 in Tab.1.

The frequency warp functions of Speaker 1 and Speaker 2 appearing in Tab.1 are showed in Fig.2. By the frequency warp function of Speaker 1 (Fig.2 (a)), all formants are shifted to higher frequency. For speaker 2, the formants below 5kHz are shifted to relatively lower frequencies, and the formants above 5kHz are shifted to relatively higher frequencies.

In order to reduce the time complexity in the search of reference points, 4 reference points, 2kHz, 4kHz, 6kHz, 7.9kHz, are used in another experiment. The experimental result for the first 4 speakers appearing in Tab.1 is showed in Tab.2. The frequency warp functions of Speaker 1 and Speaker 2 obtained by using 4 reference points are showed in Fig.3.

Tab.2. The comparison of the error rates without/with the RPA frequency warp. Four reference points are used.

No. of speaker	Error rate (Baseline)	Error rate (with RPA)	Error Reduce
1	28.2%	25.4%	9.9%
2	52.0%	39.5%	24.0%
3	47.3%	33.6%	28.9%
4	23.5%	21.0%	10.6%

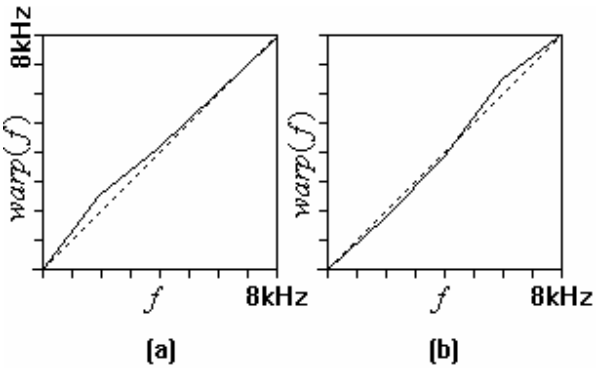


Fig.3. The frequency warp functions obtained by using 4 reference points. (a) Speaker 1 in Tab.2. (b) Speaker 2 in Tab.2.

Compared with the warping by using 8 reference points, the reference points alignment frequency warp method with 4 reference points does not achieve better performance. But speaker 3 is an exception. As mentioned above, the search process given in Section 3 can not always reach the most optimal result. The exception as for Speaker 3 is possible. By the comparison between Fig.2 and Fig.3, a speaker's frequency warp functions with different number of the reference points are similar in the trend and the variations exist only in the detail.

5. CONCLUSION

In this paper, by the examples, we explain that the variation of formant positions might be possibly inconsistent with the variation of VTLs among different speakers. For the purpose of the speaker adaptation, we give the frequency warp method that aligns the corresponding formants of different speakers by means of aligning the reference points. Although the search process can not ensure to get the optimal reference points, the experimental results show that this method takes effect. For the mixture Gaussian based HMM systems, this method can be used together with MLLR method to achieve better achievements.

6. REFERENCES

[1] Li Lee, Richard Rose, "A frequency warping approach to speaker normalization," *IEEE Transactions On Speech And Audio Processing*, Vol.6, No.1, January 1998,pp49-60.

[2] John McDonough, William Byrne, "Speaker adaptation with ALL-PASS transforms," *ICASSP*'99, paper number 2093.

[3] L. Welling, R. Haeb-Umbach, X. Aubert and N. Haberland, "A study on speaker normalization using vocal tract normalization," *ICASSP*'98, vol.2, pp.797-800.

[4] S. Umesh, L. Cohen, D. Nelson, "Frequency-warping and speaker-normalization," *ICASSP*'97, vol.2, pp.983-986.

[5] S. Umesh, L. Cohen, D. Nelson, "Improved scale-cepstral analysis in speech," *ICASSP*'98,vol.2, pp637-640.

[6] C. J. Leggetter, P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language* (1995) 9, pp171-185.

[7] Bourlard, H. And Wellekens, C.J.(1990). "Links between Markov models and multilayer perceptrons", IEEE Trans. on PAMI, vol.12, no.12, pp.1167-1178.

[8] Bourlard,H., Morgan, N.(1994) CONNECTIONIST SPEECH RECOGNITION -- A Hybrid approach, Kluwer Academic Publishers.