SPEAKER NORMALIZATION BASED ON FREQUENCY WARPING

Puming Zhan[1] *Martin Westphal*[2]

Interactive Systems Laboratories
[1]Carnegie Mellon University. Email: {zhan, ahw}@cs.cmu.edu
[2]University of Karlsruhe, Germany. Email: westphal@ira.uka.de

ABSTRACT

In speech recognition, speaker-Dependence of a speech recognition system comes from speaker-Dependence of the speech feature, and the variation of vocal tract shape is the major source of inter-speaker variations of the speech feature, though there are some other sources which also contribute. In this paper, we address the approachs of speaker normalization which aim at normalizing speaker's vocal tract length based on Frequency WarPing (FWP). The FWP is implemented in the front-end preprocessing of our speech recognition system. We investigate the formant-based and ML-based FWP in linear and nonlinear warping modes, and compare them in detail. All experimental results are based on our JANUS3 large vocabulary continuous speech recognition system and the Spanish Spontaneous Scheduling Task database (SSST).

1. INTRODUCTION

In speech recognition, we are mainly facing three major challenges: (1) speaker-dependence of the speech signal, which leads to speaker-dependence of the speech recognizer; (2) co-articulation of the speech units for acoustic models, which leads to context-dependence of the speech recognizer; (3) environmental noise, which leads to the problem of robustness of speech recognizer in practical use. Almost all kinds of speech feature are extracted from the speech signal or waveform. The reason of speaker-dependence of speech signal is very complicated. It is not only related to the physiological differences of speakers, such as vocal tract shape and fundamental pitch, but also related to the linguistic differences, such as accent, dialect and stress, etc., or even the physical and mental conditions of speakers [1]. But it is generally agreed that one of the major source of inter-speaker variance is the vocal tract shape, especially the vocal tract length (VTL) [2, 3]. Therefore, many researchers have been working on the VTL normalization via FWP in order to compensate for the speaker variation. The earlier researches were focused on the identification of isolated Vowel [1, 2, 3]. In the recent researches, the FWP was investigated in continuous speech recognition system [4, 5]. In [4], a linear FWP was investigated, and the warping factors were obtained by grid search based on Maximum-Likelihood (ML) criterion. We refer this method as ML-based FWP. The advantage of the ML-based FWP is that it guarantees to find the warping factor which is optimal in the ML criterion. The weakness of this method is that it is relatively expansive in computation. In [5], a parametric approach for FWP was proposed. We refer this method as formant-based FWP. The idea is the same as in [2, 3], i.e., the warping factors were obtained from formant estimation. But they investigated the method in large vocabulary continuous speech recognition system. The advantage of the formant-based FWP is that it is not very expansive in computation. The weakness of the method is that the warping factor is obtained only based on formant, so that it has no relationship with the ML-score and hence can not guarantee that the FWP can increase the ML-score. In this paper, we investigate the formant-based and ML-based FWP method for speaker normalization. In the formant-based FWP, instead of just using the third formant, we also investigate to use the first and second formant in our experiments. We experiment linear and nonlinear FWP in the formant-based and ML-based method, and evaluate the methods based on our JANUS3 large vocabulary continuous speech recognition system.

2. FREQUENCY WARPING

2.1. Preprocessing

The spectrum of the recorded speech signal $X(\omega)$ is assumed to be transmitted via some kind of channel and to be obtained via some kind of receiving device. In the transmitting and receiving process, the clean speech signal $S(\omega)$ is disturbed by the channel distortions and some additive noise $N(\omega)$. Most of the channel distortion $H(\omega)$ can be assumed to be multiplicative in the frequency domain leading to equation (1).

$$X(\omega) = H(\omega)S(\omega) + N(\omega) \quad (1)$$

Here we assume that $X(\omega)$ has been segmented with Hamming window, so that $H(\omega)$ and $N(\omega)$ also includes the effect of pre-emphasis and Hamming window. In the typical front-end processing of speech recognition system, $X(\omega)$ is passed through a set of Melscale filterbank which have triangular shape and is spaced in Mel scale [6, 7]. Hence the signal passing through such filterbank can be formulated as:

$$Y(i) = \sum_{\omega=\omega_{il}}^{\omega=\omega_{ih}} T_i(\omega)X(\omega) \quad 0 \le i \le N-1 \quad (2)$$

Where N is the number of filters, and ω_{il} and ω_{ih} are the lower and upper bound of the i-th filter $T_i(\omega)$. After pass-

0-8186-7919-0/97 $10.00

ing through the Melscale filterbank, the logarithm of $Y(i)$ is transformed with the DCT, so that the final N-dimensional feature vector is a set of Melscaled Frequency Cepstral Coefficients (MFCC):

$$Z(i) = \sum_{n=0}^{N-1} cos(\frac{in\pi}{N}) \log Y(n) \quad 0 \leq i \leq N-1 \quad (3)$$

Because of the logarithm in the Y-space, the multiplication of $H(\omega)S(\omega)$ in equation (1) becomes additive in the Z-space, i.e., feature-space (ignore $N(\omega)$). This is the reason that many researchers use affine transformation (rotate and/or shift Z) in the feature-space to do speaker normalization (such as Mean subtraction) and adaptation (such as MLLR). Suppose that FWP is performed in the X-space in equation (1), and the warping function is $\omega' = \varphi(\omega)$, then equation (2) becomes

$$Y'(i) = \sum_{\omega=\omega_{il}}^{\omega=\omega_{ih}} T_i(\omega)X(\varphi(\omega)) \quad 0 \leq i \leq N-1 \quad (4)$$

Comparing equation (4) to equation (2), it is clear that in most cases, the above FWP is equivalent to a nonlinear transformation in the Y-space, even in the case of linear warping, for which we assume $\varphi(\omega) = \alpha\omega$ (with constant α). Hence it is also a nonlinear transformation in the Z-space. From this point of view, considering that FWP aims to reduce the effect of frequency shift of, for example, formant positions, but not the linear channel distortions caused by the vocal tract and other speaker characteristics, it should be used together with the other affine-transform-based speaker normalization or adaptation methods.

2.2. Front-end Implementation

According to the Fourier transformation, $F(a\omega) \leftrightarrow 1/af(t/a)$, the FWP (compress or stretch in frequency axis) is equivalent to resample the waveform in time axis. Where a is the warping factor. Considering that our recognition system is synchronous in frame and all other features are extracted based on the spectrum, the FWP is implemented right after the short time spectral analysis stage in the front-end preprocessing of the system. The spectrum is warped in frequency axis frame by frame. Figure 1 is the block diagram of the JANUS front-end preprocessing. Where x_t is

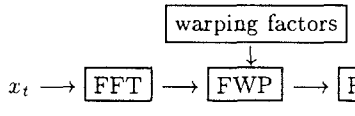

Figure 1. Diagram of FWP front-end

input speech signal, z_n is output PLP feature vector. FWP represents Frequency Warping, and PLP means Perceptual Linear Predictive [7]. The feature we are using in the experiments is the same as in [8], except here we insert a FWP step between FFT and PLP processing. The final feature is a 13-order Perceptual Linear Predictive (PLP) feature plus a power coefficient. We combine it with its delta and delta delta to form a 42-order feature vector and reduce the feature order from 42 to 28 after the LDA transformation.

We use the following piecewise warping functions in both formant-based and ML-based FWP.

Linear FWP:

$$f' = \begin{cases} \alpha_s^{-1} f & \text{if } f < \mathcal{F} \\ bf + c & \text{if } f \geq \mathcal{F} \end{cases}$$

Nonlinear FWP:

$$f' = \begin{cases} \alpha_s^{-(3f/2F_N)} f & \text{if } f < \mathcal{F} \\ bf + c & \text{if } f \geq \mathcal{F} \end{cases}$$

Where $\alpha_s = F_s/\overline{F}$ is the warping factor of speaker S, F_s is the average formant frequency of speaker S and $\overline{F}$ is the average formant frequency over all speakers in the training set. F_N is the bandwidth. b and c are constants which can be calculated according to the equations of $\alpha_s\mathcal{F} = b\mathcal{F} + c$ and $bF_N + c = F_N$. $\mathcal{F}$ is used as a threshold which aims at compensating for the bandwidth mismatch after warp. Therefore, if the frequency axis is compressed from $f = 0$ to $f = \mathcal{F}$, it will be stretched from $f = \mathcal{F}$ to $f = F_N$ in order to have $f' = f$ at the upper boundary of bandwidth. If $\mathcal{F}$ is set to F_N, then the warping is equal to those in [5].

2.3. Training Procedure

For the formant-based FWP, we use the Waves+ software to estimate formants (up to the third formant) of each speaker in the training set. The median value of each formant of each speaker and the median value of each formant over all speakers in the training set are calculated, then the warping factors α_s for every speaker are obtained. In training, we load in the warping factors and use them to warp the power spectrum (as showed in figure 1), and do the iterative training with the warped feature. For the ML-based FWP, the training principle is to find the warping factor which maximums the likelihood [4]. We use the following procedure for training:

1. Set the initial warping factor $\alpha_s = 1.0$ for all speakers
2. Do Viterbi training based on current warping factors
3. Find the best warping factor in a limited grid, that is, $\alpha_s^* = argmax_\alpha P(X_s(\varphi(f)) \mid \Lambda, W_s)$, $l_s \leq \alpha_s \leq h_s$. Where X_s is the feature vector sequence of speaker S, and W_s is the corresponding transcription. l_s and h_s are the lower and upper bound of the grid search area. They are defined as $l_s = \alpha_s - \Delta$ and $h_s = \alpha_s + \Delta$. Where α_s is the current warping factor.
4. Set $\alpha_s = \alpha_s^*$, go to step 2.

The above procedure stops if there is not significant difference in the warping factors between two consecutive training iterations.

2.4. Testing Procedure

For the ML-based FWP, we use a decoding procedure which is a little bit different from [4]. The input utterance is first decoded and aligned with the decoding output hypothesis without FWP, then the feature is warped with all possible grid points, and the ML-score is calculated with those warped features on the voiced phonemes in the path of alignment. In that case, we do not need to do forced-alignment for all warping factors. Our experimental result

shows that the error rate is almost the same as the test procedure in [4], but reduce the decoding computation. Obviously, compared to the regular test procedure, the FWP test needs to do an extra decoding and forced-alignment plus the calculation of ML score for every warping factor. For the formant-based FWP, we estimate the formants for each testing speaker in the testing set with all available testing utterances of the speaker In the test, the warping factor is used to warp the feature directly so that no warping factor search is needed as for the ML-based method.

3. EXPERIMENTS

All experiments are based on our new JANUS speech recognition system. Compared to the JANUS-II system in [8], the new system uses polyphone, instead of triphone context in the acoustic model, and clusters and splits the models based on the decision-tree. We already used the formant-based FWP on the Switchboard database and reached about 5% relative error reduction. The ML-based FWP was successfully used on the GSST (German SST) before and reduced the error rate by about 12% [9]. In the following sections we report results obtained on the SSST database comparing both methods. Compared to the database in [8], we increased about 4500 cross-talk utterances in the training set, and keep the same Devset. Thus there are 10650 utterances (5785 from 68 female speakers and 4865 from 72 male speakers), which is about 12 hours data, for training. The test vocabulary consists of 4606 words, and the language model is the class-based language model.

3.1. Distributions of the warping factors

In this section, we present the distributions (histogram statistics) of the warping factors obtained from the formant-based and ML-based FWP in the training set.

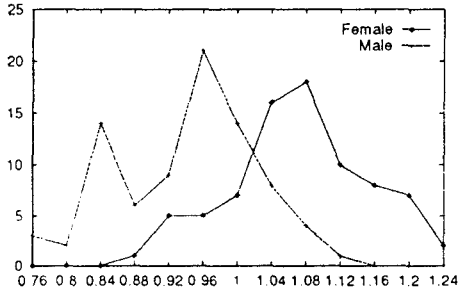

Figure 2. Histogram of F1 warping factors

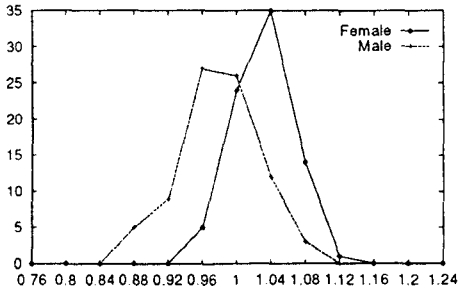

Figure 3. Histogram of F2 warping factors

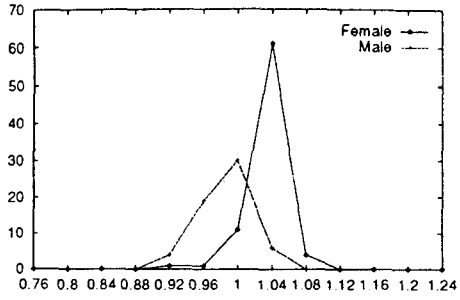

Figure 4. Histogram of F3 warping factors

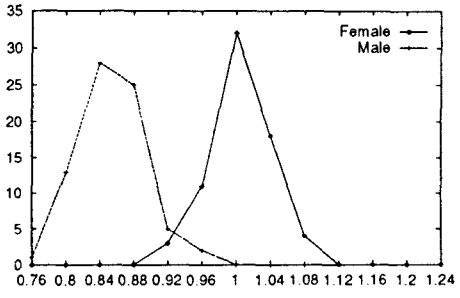

Figure 5. Histogram of ML warping factors

From the distribution figures, we find that the warping factors of female speakers are dominant in the area of $\alpha_s \geq 1.0$, and the males are dominant in the area of $\alpha_s \leq 1.0$. This coinsides with the fact that most of female's formant frequency is higher than man's. it also illustrates that the ML-based searching method can, in some extent, catch the formant variations, though it has no relation with formant estimation. But we can also find that the male and female factors in Figure 2 to 4 are not clearly separated as in 5, and they are mainly falling in the area around 1.0 and have a smaller variation (especially for the factors from the second and third formant).

3.2. Recognition results

In this section, we present the recognition results on the push-to-talk test set.

Modes	baseline	F1	F2	F3	ML
Linear	21.8%	20.5%	21.9%	21.6%	19.8%
Nonlinear	21.8%	21.5%	22.7%	21.6%	21.0%

Table 1. Word error rate of different FWPs

Table 1 contains the word error rate of the formant-based and ML-based FWP. Where F1, F2 and F3 means the formant-based FWP (first, second and third formant), and ML means ML-based FWP. It shows that (1) the linear FWP is always better than their nonlinear partner; (2) the ML-based FWP is better than the formant-based FWP. Among the formant-based FWP, we can see that F1 gave us the best result. These results are not consistent with that we used to observe from the Switchboard data, with which we observed that the nonlinear warp is better than linear warp. We also tested the cross-talk test set and observed that (1) ML-based FWP do improve the performance, but

not as much as it does on the push-to-talk test set; (2) the nonlinear FWP is still worse than the linear FWP; (3) the formant-based FWP does not improve (actually they hurt a little) the performance. From the results we can conclude that the effectiveness of FWP depends on the database, because the warping factor depends on the context, not just the speakers.

Speaker	Baseline	F1	F2	F3	ML
Meba	10.4%	9.9%	7.4%	9.9%	6.5%
Mfmm	20.5%	20.5%	23.5%	21.6%	21.6%
Mofc	11.8%	14.2%	12.8%	12.3%	11.8%
Macc	27.1%	26.0%	28.4%	27.0%	27.9%
Mrnn	31.5%	30.0%	32.1%	32.3%	27.8%
Fcba	14.0%	14.0%	16.3%	15.6%	12.1%
Fnba	15.5%	15.6%	16.5%	15.9%	14.3%
Fmcs	25.0%	21.0%	22.0%	22.9%	21.2%
Fmgl	25.0%	26.4%	26.9%	25.5%	25.5%
average	21.8%	20.5%	21.9%	21.6%	19.8%

Table 2. Word error rate for each speaker

Table 2 shows word error rate of each speaker in the test set. They were obtained based on the linear FWP. Where the first character (M/F) in speaker name represents gender. We can see that: (1) for some speakers, the FWP could not reduce their word error rate, such as Mfmm and Fmgl; (2) no warping method is consistently better than the others for all speakers, though the ML-based one is better on average; (3) it seems that the amount of error reduction is not relating to the baseline error rate. For example, for some speakers who already have a relatively low baseline error rate, such as Meba and Fcba, the FWP can still reduce the error rate. But for some speakers who have relatively high error rate, such as Mfmm and Fmgl, the FWP could not reduce their error rates. It means that FWP could not reduce the variations for some of the speakers. We think one of the reason might be the warping functions, linear or exponential, does not reflect the relationship between formants and VTL, because of the context-dependence of such relationship [10]. Another reason for the ML-based FWP might be that the warping factor, is only optimal for the models which appear in the alignment path, i.e., increasing their ML-score, it could increase more ML-score for the other models too. For the formant-based FWP, it is uncertainty if the inferior performance of is because of the formant estimation accuracy. Because Formats are context dependent, it should be better if formants are estimated based on the same context over all speakers. We calculated the average male and female VTL with the formulate in [2], and obtained 16.45cm for male (15.47cm for female) which is near the standard value (17cm) [10]. We found that F_3 is the best one for estimating VTL. This might be that vowel-dependence of the F_3 is not as strong as the first two. But we can see, from the histogram of warping factors, that the third Format does not reflect much difference among the speakers, though the average VTL value seems reasonable.

4. CONCLUSION

In this paper, we investigated the formant-based and ML-based FWP with linear and nonlinear warping functions, and reported our experimental results based on the JANUS3 large vocabulary continuous speech recognition system and the SSST database. The ML-based FWP is better than formant-based FWP. We obtained about 10% error reduction with the ML-based FWP.

5. ACKNOWLEDGMENTS

The work reported in this paper was funded in part by grants from the US Department of Defense. The author wish to thank the JANUS developers at the Interactive Systems Laboratories of University of Karlsruhe and Carnegie Mellon University for providing the JANUS system.

REFERENCES

[1] Christine Tuerk and Tony Robinson. A new frequency shift function for reducing inter-speaker variance. *EuroSpeech-93*, 1:351–354, 1993.

[2] H. Wakita. Normalization of vowels by vocal-tract length and its application to vowel identification. *IEEE Trans. ASSP*,, 25:183–192, 1977.

[3] Yoshio Ono, Hisashi Wakita, and Yunxin Zhao. Speaker normalization using constrained spectra shifts in auditory filter domain. *EuroSpeech-93*, 1:355–358, 1993.

[4] Li Lee and Richard C. Rose. Speaker normalization using efficient frequency warping procedures. *ICASSP-96*, 1:353–356, 1996.

[5] Ellen Eide and Herbert Gish. A parametric approach to vocal tract length normalization. *ICASSP-96*, 1:346–348, 1977.

[6] Charles. R Jankowski Jr., Hoang-Doan H. Vo, and Richard P. Lippmann. A comparison of signal processing front ends for automatic word recognition. *IEEE transactions on Speech and Audio Processing*, 3:286–293, 1995.

[7] Hynek Hermansky, Nelson Morgan, Aruna Bayya, and Phil Kohn. Rasta-plp speech analysis technique. *ICASSP-92*, pages I–121–124, 1992.

[8] Puming Zhan, Klaus Ries, Marsal Gavalda, Donna Gates, Alon Lavie, and Alex Waibe. Janus-ii: Towards spontaneous spanish speech recognition. *ICSLP-96*, 1996.

[9] Michael Finke, Petra Geutner, Hermann Hild, Thomas Kemp, Klaus Ries, and Martin Westphal. The karlsruhe-verbmobil speech recognition engine. *ICASSP-97*, 1997.

[10] Fant G. Speech sounds and features. 1973.