SPEAKER NORMALIZATION AND SPEAKER ADAPTATION – A COMBINATION FOR CONVERSATIONAL SPEECH RECOGNITION

Puming Zhan, Martin Westphal, Michael Finke and Alex Waibel

Interactive Systems Laboratories
Carnegie Mellon University
University of Karlsruhe, Germany
Email: {zhan, ahw}@cs.cmu.edu

ABSTRACT

Speaker normalization and speaker adaptation are two strategies to tackle the variations from speaker, channel, and environment. The vocal tract length normalization (VTLN) is an effective speaker normalization approach to compensate for the variations of vocal tract shapes. The Maximum Likelihood Linear Regression(MLLR) is a recent proposed method for speaker-adaptation. In this paper, we propose a speaker-specific Bark scale VTLN method, investigate the combination of the VTLN with MLLR, and present an iterative procedure for decoding the combined system of VTLN and MLLR. The results show that: (1) the new VTLN method is very effective with which the word error rate can be reduced up to 11%; (2) the combination of VTLN and MLLR can provide up to 15% word error reduction; (3) both VTLN and MLLR are more effective for the push-to-talk data than for the cross-talk data.

1 INTRODUCTION

Almost all speech recognizers are, in some extent, sensitive to the variations of speakers and/or environment. The performance of a current state-of-the-art speech recognition system could vary largely in practical use because of these variations. The speaker-dependent speech recognition system comes from the speaker-dependent speech signal. The reason that the speech signal is speaker-dependent is very complex. It is not only related to the physiological differences of speakers, but also related to the linguistic differences [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, some researchers have been devoted to the vocal tract length normalization (VTLN) for speaker normalization [1, 2, 3, 4, 5, 6]. Generally speaking, two issues are involved in VTLN: (1) Given the speech data from a speaker, how to obtain the warping factor for normalization; (2) Given a warping factor, how to do the normalization; The warping factors could be obtained via formant calculation as in [2, 3, 5], or via line search as in [4, 6]. The normalization could be implemented in the Fourier spectrum domain as in [2, 5, 6], or in the Bark domain as in [3, 4]. We used the VTLN based on Fourier spectrum warping in [7], and estimated the warping factor via formant calculation or line search. We obtained up to 10% word error reduction with the line searching warping factor, and did not get any improvement with the formant method. The Fourier spectrum warping VTLN has some disadvantages, such as, exists the bandwidth mismatch, the need to specify the warping rule, the need to interpolate the warped spectrum, etc. Therefore, we propose a speaker-specific Bark scale VTLN in this paper, with which those disadvantages can be eliminated.

However, speaker variation is only one of the major variation source. There are vast unpredictable channel and environmental variations that the speech recognizers have to face with in practical use. Speaker adaptation is a technique, with which a speech recognizer can be adapted towards a new speaker and/or environment with a small amount of adaptation data, or even without adaptation data (unsupervised adaptation). The MLLR adaptation linearly transforms a speaker-independent (SI) system towards a speaker-dependent system in the acoustic model space based on adaptation data [8, 9]. As we will show in this paper that the VTLN is equivalent to a nonlinear transformation of the speech signal in the feature space. Hence it is interesting to investigate the combination of the two methods in a speech recognition system. Intuitively, speaker-normalization could be helpful for speaker-adaptation to learn the new speaker and/or environment faster in a limited adaptation data, because the normalized speech features are less variant than the original one.

In this paper, we propose the speaker-specific Bark scale VTLN, which can be implemented in the front-end of a speech recognition system. Then we investigate the combination of the VTLN and MLLR, and propose an iterative test procedure for decoding the combined system of VTLN and MLLR. We also compare results for the push-to-talk and cross-talk speech data. All experimental results are obtained from our JANUS-III large vocabulary continuous speech recognition system based on the SSST database.

2 VTLN IN THE BARK DOMAIN

2.1 Preprocessing

The recorded speech signal is assumed to be transmitted via some kind of channel and to be received via some kind of receiving device. In the transmitting and receiving process, the clean speech signal is disturbed by the channel distortions and some additive noises. Generally, the channel distortion is assumed to be multiplicative in the frequency domain, so that the received speech signal can be expressed as equation (1):

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

Where $X(\omega)$, $S(\omega)$, $H(\omega)$, and $N(\omega)$ are the spectrum of the received speech signal, the clean speech signal, the channel response, and the additive noise signal. We assume that $X(\omega)$ has been segmented with a Hamming window, so that $H(\omega)$ and $N(\omega)$ also include the effect of pre-emphasis and the Hamming window. In the Bark filter bank front-end, $X(\omega)$ is integrated with the filter bank using band pass filters spaced according to the Bark scale, and usually have triangular or trapezoid shape. The integration with the filter bank can be formulated as:

$$Y(n) = \sum_{\omega=l_n}^{\omega=h_n} T_n(\omega)X(\omega) \quad 0 \le n \le N-1 \quad (2)$$

Where $Y(n)$ is n-th filter bank coefficient, N is the number of filters, l_n and h_n are the lower and upper bound of the n-th filter $T_n(\omega)$. The bandwidth of each $T_n(\omega)$, i.e., $h_n - l_n$, depends on the Bark scale.

2.2 VTLN based on speaker-specific Bark scale

We view the measured Bark scale presented in [10, 11] as the average scale which applies to all speakers. However, for a specific speaker, the Bark scale should be different in some extent due to the specific vocal tract length/shape. Our approach to do VTLN in Bark domain is not directly to adjust the filter bank space or to shift Bark coefficient as in [3, 4]. Instead, we find a specific Bark scale for each speaker, and use this speaker-specific Bark scale to compress the speaker's spectrum. The VTLN is implemented in the process of filter bank integration under the speaker-specific Bark scale. We refer this method as speaker-specific Bark scale warping. Figure 1 is the block diagram of the speaker-specific Bark scale front-end.

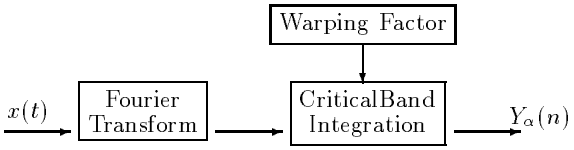

Figure 1: Speaker-specific Bark/Mel scale VTLN

Compared to the Fourier spectrum warping as in [7], the VTLN is implemented in the Bark domain by the speaker-specific Bark scale filter bank integration. Figure 1 can be expressed as equation (3):

$$Y_\alpha(n) = \sum_{\omega=l_\alpha(n)}^{\omega=h_\alpha(n)} T_n(\omega)X(\omega) \quad 0 \le n \le N-1 \quad (3)$$

Compared to equation (2), the difference is that the filter bank space, i.e., $h_\alpha(n) - l_\alpha(n)$, depends on the speaker-specific warping factor α, because each speaker has a specific Bark scale. We define the speaker-specific Bark scale as equation (4):

$$B_\alpha(\omega) = 6ln(\omega/(1200\pi\alpha) + \sqrt{(\omega/(1200\pi\alpha))^2 + 1)}) \quad (4)$$

Where α is the speaker-specific parameter. If we let $\alpha = 1.0$ for all speakers, equation (4) becomes equation (3) in [10]. Figure 2 is the warping curves of the speaker-specific Bark scale.

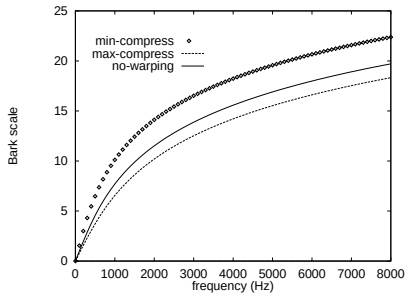

Figure 2: Bark scale warping curves

Three curves are presented in figure 2, which reflect the range of the warping factors obtained during training. The lower and upper curves correspond to the minimum and maximum factors, and the middle one corresponds to unit warping factor (no warping). The area between the upper and lower curve are the possible range of warping factors obtained in our training set. We observed that the warping factors of female speakers are dominant in the area between the lower and middle curve, which corresponds to more spectrum compress, and the warping factors of male speakers are dominant in the area between the middle and upper curve, which corresponds to less spectrum compress. This is consistent with the fact that female's VTL is generally shorter than male's, and the formant positions are higher than the male's in frequency axis. Thus for the normalization purpose, in general, most of the female's spectrum should get more compress towards the standard one, and vice visa for male's spectrum. The result is also consistent with what we obtained in [7], though the normalization method is different. The major advantage of the speaker-specific Bark scale VTLN is that it is very simple and the performance is also better. Compared to the VTLN in frequency domain and one-Bark-shift method in [3], there is no warping rule need to be specified, no spectrum interpolation need to be handled, and no bandwidth mismatch problem. We use the same training procedure as in [7] to train the VTLN system.

3 COMPARISON OF VTLN AND MLLR

Let $\mathbf{z}_\alpha(\mathbf{t})$ be a N-dimension feature vector sequence which is derived from $Y_\alpha(n)$ in equation (3) (usually cepstrum of $Y_\alpha(n)$), and be used to train the SI system. Let $\mathbf{o}_\alpha(\mathbf{t})$ be a N-dimensional feature vector sequence from a new speaker. The mixture Gaussian density for $\mathbf{z}_\alpha(\mathbf{t})$ at state i in the HMM can be expressed as:

$$P(\mathbf{z}_\alpha(\mathbf{t}) \mid i) = \sum_k p(w_k \mid i) N(\mathbf{z}_\alpha(\mathbf{t}); \mu_{ik}, \Sigma_{ik}) \quad (5)$$

Where $p(w_k \mid i)$ is the probability of the kth mixture component of state i, $N(\mathbf{z}_\alpha(\mathbf{t}); \mu_{ik}, \Sigma_{ik})$ is kth single Gaussian density with mean μ_{ik} and covariance matrix Σ_{ik}.

In the MLLR adaptation, it is assumed that $\mathbf{o}_\alpha(\mathbf{t})$ has linear relationship with $\mathbf{z}_\alpha(\mathbf{t})$ as $\mathbf{o}_\alpha(\mathbf{t}) = A_i\mathbf{z}_\alpha(\mathbf{t}) + b_i$ in [8]. Where A_i is a $N \times N$ matrix and b_i is a $N \times 1$ vector in state i. Therefore, the probability density of observation $\mathbf{o}_\alpha(\mathbf{t})$ is obtained by replacing the single Gaussian density with $N(\mathbf{o}_\alpha(\mathbf{t})); A_i\mu_{ik} + b_i, A_i\Sigma_{ik}A_i^T)$ in equation (5). The adaptation algorithm is to estimate A_i and b_i

to maximum $P(\mathbf{o}_\alpha(\mathbf{t}) \mid i)$. A_i was assumed to be a diagonal matrix to avoid the expansive computation in [8]. This assumption was eliminated by just linearly transform the mean vector in [9], i.e., replacing the single Gaussian density with $N(\mathbf{o}_\alpha(\mathbf{t}); A_i\mu_{ik} + b_i, \Sigma_{ik})$ in equation (5). This is equivalent to assume $\mathbf{o}_\alpha(\mathbf{t}) = \mathbf{z}_\alpha(\mathbf{t}) + \tilde{b_i}$, where $\tilde{b_i} = (A_i - I)\mu_{ik} + b_i$, and I is a unit matrix.

We use α in $\mathbf{z}_\alpha(\mathbf{t})$ and $\mathbf{o}_\alpha(\mathbf{t})$ to illustrate that the speech feature we are using for MLLR adaptation is the normalized feature with VTLN. From equation (3) we can see that the relationship between α and $Y_\alpha(n)$ is generally nonlinear. Thus the VTLN can not be completely merged into the MLLR linear adaptation. Therefore, the combination of VTLN and MLLR should present a further global improvement for the system. Actually, VTLN could help MLLR in two ways. In the unsupervised adaptation mode, the system with VTLN can give a better hypothesis to guide the MLLR adaptation. In the supervised adaptation mode, VTLN could reduce the variations of the adaptation data, and make the very limited data more effective for estimating the transformation matrices.

4 DECODING PROCEDURE FOR THE COMBINED SYSTEM

In this section, we propose a iterative procedure for decoding the combined system of VTLN and MLLR. Suppose that the standard system is a SI system which was trained with the VTLN speech feature. We use the MLLR in the on-line unsupervised mode. Thus, for a given utterance, we first need to find the best warping factor based on the training criterion (ML), then decode with the warped feature to obtain the hypothesis, estimate the MLLR transformation matrices based on the hypothesis and transform the model parameters, and finally decode again with the transformed model. We can run the whole procedure iteratively on one utterance to increase the ML score and hopefully increase the word accuracy. Following is the iterative decoding procedure:

1. Set the initial warping factor $\alpha = 1.0$, and load the SI system model parameters Λ.
2. Decode the input utterance.
   $\hat{W} = \arg\max_W P(W \mid \mathbf{o}_\alpha(\mathbf{t}), \Lambda)$
3. Do Viterbi alignment to get the best state segment.
   $s_t^* = \arg\max_{s_t} P(o_\alpha(t), s_t \mid \Lambda, \hat{W})$
4. Find the best warping factor based on the segment.
   $\alpha^* = \arg\max_\alpha P(\mathbf{o}_\alpha(\mathbf{t}) \mid s_t^*, \Lambda)$
5. Decode based on the best warping factor.
   $\tilde{W} = \arg\max_W P(W \mid \mathbf{o}_{\alpha^*}(\mathbf{t}), \Lambda)$
6. Calculate the MLLR transformation matrices T.
   $T^* = \arg\max_T P(\mathbf{o}_{\alpha^*}(\mathbf{t}) \mid T(\Lambda), \tilde{W})$
7. decode again with the transformed models.
   $\check{W} = \arg\max_W P(W \mid \mathbf{o}_{\alpha^*}(\mathbf{t}), T^*(\Lambda))$
8. Let $\hat{W} = \check{W}$, $\alpha = \alpha^*$, $\Lambda = T^*(\Lambda)$, and go to step 3.

The above procedure stops if there is no significant increase of the ML score between two consecutive iterations. The step (1) $\longrightarrow$ (5) is the decoding procedure for a VTLN system, and step (1) $\longrightarrow$ (2) $\longrightarrow$ (6) $\longrightarrow$ (7) is the decoding procedure of a system with unsupervised MLLR adaptation (replacing $\tilde{W}$ with $\hat{W}$ in step (6)). Compared to the regular decoding method, which only need to run step (2) for each utterance, the above iterative decoding is very expansive. It needs to run twice as long as the regular decoding procedure in each iteration, and some extra computation for the best warping factor and the MLLR transformation matrices.

5 EXPERIMENTS

All experiments are based on our JANUS-III speech recognition system. The SSST database composed of 1/3 push-to-talk dialogs and 2/3 cross-talk dialogs. We use the same database as in [7]. Readers can find detail analysis of push-to-talk and cross-talk data in [12]. We use the push-to-talk and cross-talk dialogs together to train the acoustic models, but keep an individual test set for each of them. The push-to-talk test set consists of 86 utterances, the cross-talk test set consists of 117 utterances. The test vocabulary consists of 4606 words. The out of vocabulary word rate is 2.35% for push-to-talk test set, and 0.89% for cross-talk test set. The language model is the class-based trigram language model.

We use the same Perceptual Linear Predictive (PLP) cepstral coefficients as in [10], except the bark scale is speaker-specific as equation (4). We calculate 21 filter bank coefficients and use them to derive 13 LPC-Driven cepstral coefficients. Then the cepstral coefficients and power are combined with their first and second derivative to generate a 42-dimensional feature vector. Finally, this vector is transformed with the linear discriminant analysis (LDA) matrix, and reduced to 28 coefficients.

5.1 Results of the combined system

We test MLLR in the unsupervised mode, and assume that only the current input utterance is available for the estimation of the transformation matrices. We run one iteration of the decoding procedure for the combined system. The results are obtained on the push-to-talk test set.

Spk	SI	MLLR	VTLN	VtlnMllr
Meba	10.4%	4.7/7.3%	10.4/8.6%	5.6/6.9%
Mfmm	20.5%	16.7/20.5%	19.3/21.6%	13.4/20.1%
Mofc	11.8%	8.0/11.8%	9.4/8.5%	5.2/8.5%
Macc	27.1%	22.5/27.7%	26.5/26.1%	21.3/25.9%
Mrnn	31.5%	18.8/30.2%	26.5/28.7%	18.2/28.5%
Fcba	14.0%	12.1/16.7%	16.7/14.4%	10.7/13.9%
Fnba	15.5%	10.4/14.9%	12.3/13.3%	10.4/13.3%
Fmcs	25.0%	16.4/23.1%	21.6/22.1%	16.0/21.4%
Fmgl	25.0%	20.4/27.4%	22.4/22.5%	13.2/22.5%
AVE	21.8%	15.3/21.3%	19.1/19.4%	14.0/18.6%

Table 1: WER of VTLN, MLLR and VTLN+MLLR

In Table 1, the SI column shows the word error rate (WER) of the SI system, and VtlnMllr column shows the WER of the combined system of MLLR and VTLN. We give two WERs for each testing speaker to represent the WER obtained with transcription/hypothesis as guide for the warp factor and transformation matrices estimation. Therefore, We can observe the real co-effect of VTLN and MLLR without the effect of the recognition errors from the transcription based results. We can also observe the sensitiveness of VTLN and MLLR to the recognition errors from the hypothesis based results. From Table 1 we can see: (1). the speaker-specific Bark scale VTLN can

reduce up to 11% word errors; (2). MLLR is very sensitive to the recognition errors, the average WER increases from 15.3% to 21.3% when using the hypothesis, instead of the transcription, to guide the estimation of transformation matrices. For example, for speaker Meba, the WER increased from 4.7% to 7.3% when using the hypothesis, though the baseline WER is very low for this speaker. (3). the on-line unsupervised MLLR is not very effective in the case that the baseline system has about 20% WER. (4). VTLN is not sensitive to the recognition errors, and it can significantly improve the recognition accuracy; (5) the combination of MLLR and VTLN can improve the performance further, though the MLLR eats a small part of the gains from VTLN.

5.2 Results of the iterative decoding

In this section, we present the results of the iterative decoding procedure described in section 4. We run three iterations of the procedure for the combined system.

Baseline	Iter.1	Iter.2	Iter.3
21.8%	18.6%	18.4%	18.4%

Table 2: WER of the iterative decoding

Table 2 shows that there is a slight improvement from iteration 1 to iteration 2, and we also observed the increase of the ML score. But after two iterations, it seems that there is only a minor ML score increase, and we did not observe the improvement of WER. This means that the error reduction of the hypothesis in each iteration is not enough to give a good guide for the estimation of warping factor and transformation matrices for the next iteration.

5.3 Comparison of push-to-talk and cross-talk

In this section, we present some testing results on the cross-talk test set for comparison with the push-to-talk data.

Baseline	MLLR	VTLN	VtlnMllr
23.4%	16.7/24.8%	22.0/22.5%	15.8/24.2%

Table 3: WER of cross-talk test set

Table 3 shows that the VTLN can improve the WER, but not as effective as it does for the push-to-talk data, and the MLLR does not help. One of the reason is that the average length of the cross-talk utterances is only 9.5 words per utterance (compared to 38.5 words of the push-to-talk utterances) This could be a problem, since our VTLN and MLLR only use the current utterance to estimate the warping factor and the transformation matrices. In addition, the two male speakers in cross-talk test set have very high WER (about 50%). This could affect the VTLN and MLLR, since the warping factor and the transformation matrices were estimated based on the poor hypothesis. Again, MLLR is very sensitive to the recognition errors compared to VTLN. But we also found from Table 3 that the VTLN and MLLR are still effective for the cross-talk data if the transcription is used to guide the estimation of warping factor and matrices, though the utterances are very short.

6 CONCLUSION

In this paper, we proposed a speaker-specific Bark scale VTLN method, investigated the combination of VTLN and MLLR, and present an iterative procedure for decoding the combined system of VTLN and MLLR. The new VTLN reduced up to 11% word errors, and the combination of VTLN and MLLR can reduce word errors up to 15%. We also found that MLLR is very sensitive to the recognition errors in the case of unsupervised adaptation, and VTLN is not. Both are more effective for the push-to-talk data than the cross-talk data.

7 ACKNOWLEDGMENTS

The work reported in this paper was funded in part by grants from the US Department of Defense.

8 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, 1996.

[6] Steven Wegmann, Don McAllaster, Jeremy Orloff, and Barbara Peskin. Speaker normalization on conversational telephone speech. *ICASSP-96*, 1:339–341, 1996.

[7] Puming Zhan and Martin Westphal. Speaker normalization based on frequency warping. *Proceedings of ICASSP97, Munich, Germany*, 1997.

[8] Vassilios V. Digalakis, Dimitry Rtischev, and Leonardo G. Neumeyer. Speaker adaptation using constrained estimation of gaussian mixtures. *IEEE Trans. Speech and Audio Processing*, 3:357–365, 1995.

[9] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hmms. *Computer Speech and Language*, 9:171–186, 1995.

[10] Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *J. Acoust. Soc. Am. 87(4)*, pages 1738–1752, 1990.

[11] E. Zwicker and E. Terhardt. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *J. Acoust. Soc. Am. 68(5)*, pages 1523–1525, 1980.

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