REGION-BASED VOCAL TRACT LENGTH NORMALIZATION FOR ASR

Michail G. Maragakis, Alexandros Potamianos

Technical University of Crete, Dept. of Electronics and Computer Engineering,
Chania 73100, Greece

{potam, melidoni}@telecom.tuc.gr

ABSTRACT

In this paper, we propose a Region-based multi-parametric Vocal Tract Length Normalization (R-VTLN) algorithm for the problem of automatic speech recognition (ASR). The proposed algorithm extends the well-established mono-parametric utterance-based VTLN algorithm of Lee and Rose [1] by dividing the speech frames of a test utterance into regions and by warping independently the features corresponding to each region using a maximum likelihood criterion. We propose two algorithms for classifying frames into regions: (i) an unsupervised clustering algorithm based on spectral distance, and (ii) an unsupervised algorithm assigning frames to regions based on phonetic-class labels obtained from the first recognition pass. We also investigate the ability of various mono-parametric and multi-parametric warping functions to reduce the spectral distance between two speakers, as a function of phone. R-VTLN is shown to significantly outperform mono-parametric VTLN in terms of word accuracy for the AURORA4 database.

Index Terms— Speech recognition, Acoustic signal processing, Speech processing.

1. INTRODUCTION

VTLN is a well-established speaker normalization algorithm used to improve automatic speech recognition (ASR) performance. VTLN compensates for the effect of speaker dependent vocal tract lengths by warping the frequency axis of the spectrum magnitude before computing the cepstrum coefficients [1]. VTLN is applied in the feature space using warping functions that typically depend only on a few free parameters [2]. Even with a single free parameter (the warping factor α) and using very few data for estimation (typically a single utterance), VTLN performs well for a variety of recognition tasks. This unique free parameter can be obtained by calculating formants frequencies [3] or by using a maximum likelihood (ML) criterion usually in a two-pass speech recognition scenario [1, 4, 5].

Lee et al. [1] proposed an efficient maximum likelihood algorithm for estimating the warping factor for linear frequency scaling. In [5], VTLN was used during both the training and the testing phase. In these approaches, a single warping factor and function is used for each utterance; this function may be linear, piecewise linear or non-linear, e.g., power [6]. It is well known from the speech analysis literature, that spectral differences among speakers due to varying vocal tract length are both phone-dependent and non-linear and cannot be fully captured by a single warping function and factor selected on a per utterance basis. Recently, there have been attempts to compute "instantaneous" warping factors, i.e., warping factors on a per frame basis. Among these frame-based VTLN approaches the most notable are the MATE framework [7] where spectral warping is applied to individual frames using a two-dimensional Viterbi decoding algorithm to estimate the frame-based warping factor and [8] where the best warping factor is selected based on a normalized codebook.

In this study, the dependence between warping and phones is investigated. Based on the conclusions drawn from this analysis, algorithms for the division of test utterance' s frames into regions and the estimation of an optimal, for each region, warping factor and function is provided. The two-pass recognition method of [1] is extended so that region-dependent optimal warping factor and function can be obtained from a set of candidates, based on an ML criterion. The proposed region-based VTLN (R-VTLN) algorithm adds little computational complexity (since factors can be independently estimated in each region) and captures most of the improvement of frame-based VTLN (over utterance-based VTLN).

The paper is organized as follows. In Section 2, the frequency warping based speaker normalization procedure is described. In Section 3, we examine the ability of various warping functions to reduce the spectral distance between speakers for different phones. The Region-based VTLN (R-VTLN) algorithm is proposed in Section 4. In Section 5, results are presented comparing R-VTLN and VTLN and the paper concludes in Section 6.

2. SPEAKER NORMALIZATION USING MONO-PARAMETRIC FREQUENCY WARPING

According to [1], for each utterance an optimal warping factor $\hat{\alpha}$ is selected from a discrete ensemble of M possible values so that the likelihood of the warped utterance is maximized with respect to a given speech recognition model (ensemble of hidden Markov models) and a given transcription. The transcription is obtained from a first recognition pass. After the evaluation of the optimal factor, a warped process,

$$\omega \rightarrow \tilde{\omega} = g(\alpha) \cdot \omega, \tag{1}$$

is obtained, where ω and $\tilde{\omega}$ are correspondingly the unwarped and warped frequencies and g is the warping function.

Let $X^{\alpha} = g_{\alpha}(X)$ denote the sequence of cepstral vectors where each one of them is warped by the warping function $g_{\alpha}()$. If λ denotes the parameters of the HMM models and W is the transcription obtained from an initial recognition pass, the optimal warping factor (referred as *global* henceforth) is defined as

$$\hat{\alpha}_{glb} = \arg\max_{\alpha} P(X^{\alpha}|\lambda, W) \tag{2}$$

After the estimation of the $\hat{\alpha}_{glb}$, the frequency warped observation vector $X^{\hat{\alpha}_{glb}}$ is decoded in a second recognition pass to obtain the decoded transcription.

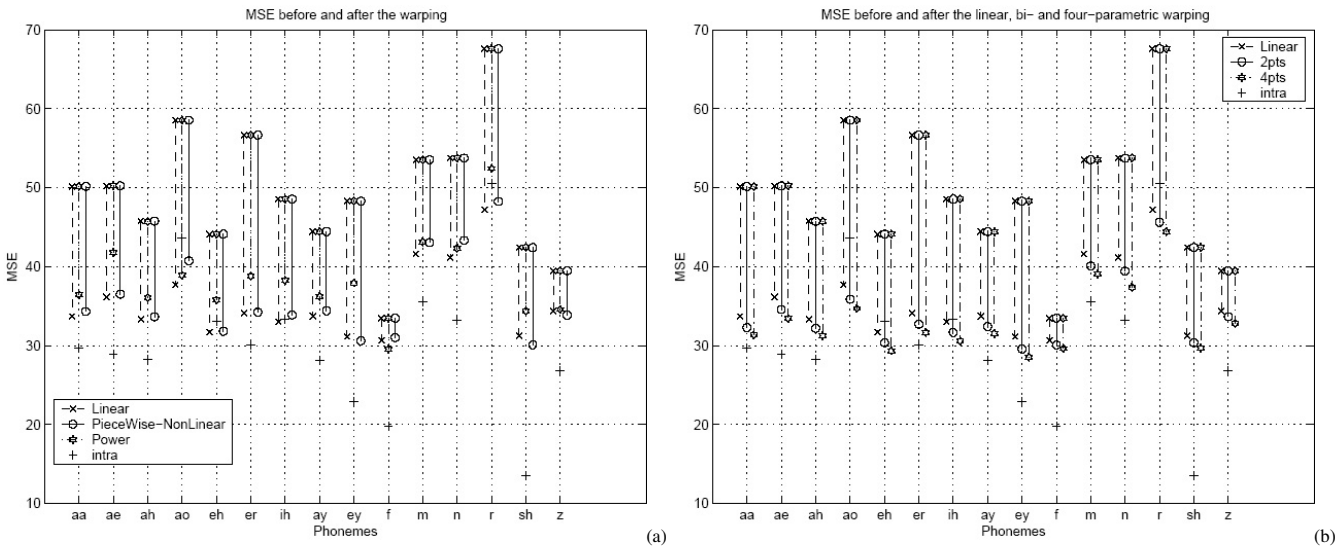

Fig. 1. Intra-speaker variability (+) and averaged MSE between reference and mapped speakers (male and female) before and after warping: (a) linear, piecewise-nonlinear and power warping functions (b) bi-parametric (2pts) and four-parametric (4pts) warping.

3. ANALYSIS OF WARPING FUNCTIONS

In this section, we evaluate the ability of various warping functions $g_\alpha()$ to reduce the spectral mismatch between speakers. A variety of mono-parametric and multi-parametric warping functions are evaluated. The reduction of the spectral distance between two instances of the same phone by different speakers is computed using a Minimum Square Error (MSE) criterion. For this purpose, a subset of the TIMIT training dataset is used comprising of 16 speakers (8 male and 8 female). Speakers are separated into "reference" and "mapped" speakers (that get "warped" to the reference speakers). The MSE criterion is defined as follows: Given the unwarped spectrum Y (reference spectral envelope computed using a mel-based ASR front-end), the warped spectrum Z (mapped spectrum) and the warping function g_α, the MSE is defined as,

$$MSE = \frac{1}{K}\sum_{i=1}^{K}\Big(Y_i - g_\alpha(Z_i)\Big)^2 \tag{3}$$

where K=256, the number of mel spectral coefficients. The frequency warping is performed by taking the middle frame of two instances of a phone (steady-state), computing the smooth spectral envelope, and then the optimal warping factor $\hat{\alpha}$ is computed, so that the MSE between the warped spectrum $g_\alpha(Z)$ and the corresponding unwarped spectrum, Y is minimized, i.e.

$$\hat{\alpha} = argmin_\alpha \frac{1}{N}\sum_{i=1}^{N}\Big(Y_i - g_\alpha(Z_i)\Big)^2 \tag{4}$$

Optimization is achieved by a full search in the interval of warping factors ranging from 0.88 to 1.12, where 1 corresponds to no warping as in [1] and similarly minimizing the MSE over all parameters for multi-parametric warping as described next. Frequency warping is implemented by re-sampling the smooth spectral envelope according to the warping factor $\hat{\alpha}$.

The following warping function have been investigated: (i) linear, (ii) power [6], (iii) piecewise non-linear [9], (iv) bi-parametric piece-wise linear and (v) four-parametric piece-wise linear. For the bi-parametric case different factors (α_L and α_H) for the low ($f < 3$ kHz) and high ($f \geq 3$ kHz) frequencies correspondingly, are obtained. An optimal global warping factor $\hat{\alpha}$ is computed and the two parameters are computed under the constraint $|\alpha_L - \hat{\alpha}| \leq 0.04|$ and $|\alpha_H - \hat{\alpha}| \leq 0.04|$. For the four-parametric case the warping factors correspond to the frequency ranges 0-1500 Hz, 1500-3000 Hz, 3000-4500 Hz and 4500-8000 Hz.

Figure 1(a),(b) show average MSE reduction due to warping for the warping functions presented above. The linear and piecewise nonlinear warping function are shown to perform somewhat better than the power function in (a). Note the relatively large distance reduction for vowels, glides and the small reduction for fricatives /f/ and /z/, demonstrating the dependence between frequency warping functions and phones, especially for the vowels. Note that the average warping factors (amount of scaling) are higher for vowels than for fricatives (not shown here). In (b) further reduction of the spectral distance is provided by using two and four-parametric warping functions. However, the additional improvement over linear warping was not as large as expected. Finally, note that the average MSE after warping is similar to the intra speaker variability MSE threshold (distance between two repetition of the same speaker – shown as a cross) for some phones.

4. REGION BASED VTLN

In this section, we present the Region-based VTLN algorithm (R-VTLN) that first categorizes the testing utterance's frames into regions and then region-specific spectral warping functions and factors are computed using an ML criterion in order to optimally warp each region's testing frames.

Two algorithms are proposed for classifying frames into regions. Specifically each utterance's cepstral vectors are classified through either

- an unsupervised KMeans algorithm (referred as KM henceforth) based on cepstral distance, or
- an unsupervised,based on the conclusions from Section 3, algorithm, assigning frames to regions based on phonetic-class

labels obtained from the first recognition pass (referred as *PhCat* henceforth).

Their output is a mapping F between the L frames and their region index sequence R, $F : l \rightarrow p$. Following the frame classification algorithm, median filtering is applied on the sequence R. Median filtering is used to smooth these inherently noisy frame assignments, based on the continuity criterion.

After the categorization, the spectral coefficients corresponding to each region are warped according to one of the M factors α and one of the N functions g. This results to a multi-dimensional warping process. The multi-dimensional warping process obtains the P optimal factors and functions for each region

$$\vec{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ ... \\ \alpha_P \end{bmatrix}, \vec{g} = \begin{bmatrix} g_1 \\ g_2 \\ ... \\ g_P \end{bmatrix}$$

by maximizing the likelihood of the warped vectors with respect to the transcriptions from the first pass W and the unnormalized HMM λ,

$$\hat{\vec{\alpha}}, \hat{\vec{g}} = argmax_{\vec{\alpha},\vec{g}} P(X^{\vec{\alpha},\vec{g}} | \lambda, W) \quad (5)$$

The optimal parameters can be determined by

- an exhaustive search over factors and functions for all the regions simultaneously (referred as *Sim* henceforth) or
- by searching independently for each region, while the other regions are warped linearly by $\hat{\alpha}_{glb}$ (referred as *Sep* henceforth).

To summarize during recognition the following two pass strategy is followed:

- through a first recognition pass, a transcription W is obtained using the unwarped sequence of cepstral vectors X and the unnormalized model λ.
- The utterance's frames are categorized into P regions.
- For each region, an optimal warping factor and function is evaluated,
- The warped with $\hat{\vec{\alpha}}$ and $\hat{\vec{g}}$ sequence $X^{\hat{\vec{\alpha}},\hat{\vec{g}}}$ is decoded in order to obtain the final recognition result.

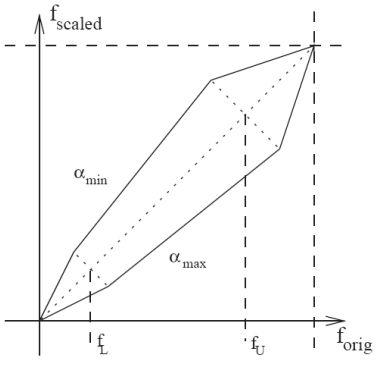

Fig. 2. Linear and Piecewise Linear warping functions.

5. EXPERIMENTAL SETTING AND RESULTS

This section presents the experimental setup and the evaluation results for the R-VTLN algorithm presented in the previous section. The training set we used was the AURORA4 clean set (7138 utterances, 128.294 words). The test consists of the AURORA4's clean test set (330 utterances from 8 speakers, 5353 words). HMM monophone models with three states per phone and one, three and eight Gaussians per state were trained. Feature extraction consisted of a Hamming window 25 ms and a frame update of 10 ms, resulting in the standard 39 dimensional cepstrum coefficients (MFCC). The same front-end was used also to compute spectral distances during the frame classification procedure (KM). For the PhCat classification algorithm, regions were selected based on the average warping factors of phonemes from our analysis. For the two-region case, the split was between vowel and diphthongs vs. the rest. For the three-region case, silence is excluded from the second class and a third class is created just for silence. For the five-region case the regions are as follows: (1) /ey/, /ay/, /aa/, /ae/, /iy/, /ih/, /ah/, (2) /uh/, /uw/, /aw/, /ao/, /ow/, /oy/, /er/, /eh/, (3) /jh/, /ch/, /dh/, /sh/, /th/, /zh/, /m/, /n/, /hh/ ,/f/ (4) /g/, /k/, /w/, /d/, /b/, /p/, /t/, and (5) /l/, /z/, /j/, /s/, /sh/, /r/ and silence. For all experiments the optimum, for each region, warping factor is obtained by searching between $0.88 \leq \alpha_m \leq 1.12$ with step 0.02. The warping functions evaluated are either Linear and Piecewise-Linear as shown in Fig. 2.

Regions	*2*	*3*	*5*
Baseline	48.3		
VTLN (two-pass	53.2		
R-VTLN KM-Sim	56.2	-	-
R-VTLN PhCat-Sim	55.6	-	-
R-VTLN KM-Sep	56.1	55.7	55.4
R-VTLN PhCat-Sep	55.8	55.6	55.4

Table 1. Word accuracy results (%) evaluated on clean test set of AURORA4.

Two Regions Case			
GMM per State	*1*	*3*	*8*
Baseline	48.3	55.4	56.4
VTLN (two-pass)	53.2	60.1	60.7
R-VTLN KM-Sep	56.1	63.5	63.9

Table 2. Word accuracy results (%) versus the number of Gaussian Mixtures per State on monophones HMM.

The experimental results of the proposed R-VTLN are shown at Table 1. At the same Table, for comparison reasons, the baseline and the results obtained by the VTLN algorithm of Lee and Rose [1] are also presented. Results for all four R-VTLN variant are shown, namely KM and PhCat referring to the frame classification method, and Sim and Sep depending on the warping factor estimation method. The word accuracy (%) is shown for the AURORA 4 clean test set.[1] We observe that VTLN significantly improves on the baseline and that all variants of R-VTLN outperform

[1]Given the small difference in performance between the Sim and Sep algorithms, only results for the Sep algorithms are shown for three and five regions.

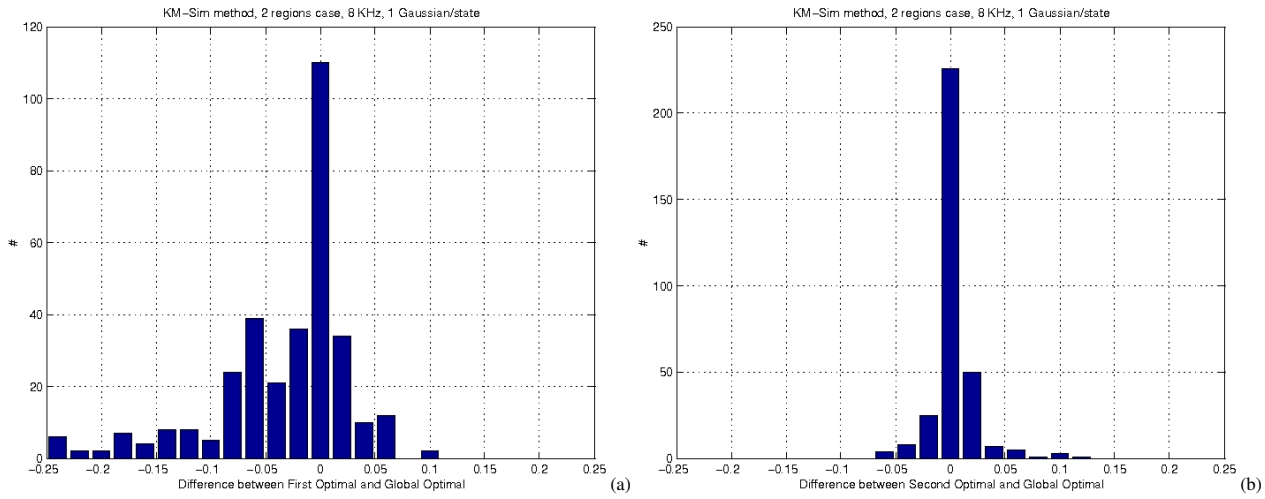

Fig. 3. Distribution of the distance between the obtained by *KM-Sim* optimal factors and the global factor $\hat{\alpha}_{glb}$: (a) For the first region, (b) For the second region.

utterance-based VTLN over all conditions. There is no significant difference between the variants of R-VTLN, thus the simplest and most computationally efficient KM-Sep algorithm is used for further experimentation. Also the improvement when increasing the number of regions from two to three is not significant. Results degrade somewhat when using five regions. This could be due to the lack of adequate data to estimate multiple parameters (a single utterance is used here for warping factors and functions estimation) and decreasing returns from using multiple regions.

Next we investigate if these improvements hold for HMM models of increasing complexity. Results are presented in terms of word accuracy for the two regions case and the KM-Sep method for the AURORA 4 task at Table 2, for one, three and eight Gaussians per state. As expected baseline performance increases significantly. At the same time the relative improvement of VTLN over baseline decreases. However, the improvement of R-VTLN over VTLN remains consistently the same.

Finally, in Fig. 3 the distributions of the difference between the region-based warping factors and the global warping factor $\hat{\alpha}_{glb}$ are shown for the KM-Sim method. As expected the region factors lie around the global optimal factor, and take lower values on average for the first region and somewhat higher for the second region.

6. CONCLUSIONS

In this paper, we have shown quantitatively the dependence between frequency warping functions and phones. Based on these results we proposed a region-based VTLN algorithm where, first, frames are classified in regions and then a region-dependent warping is applied based on an ML criterion. R-VTLN was evaluated on AURORA4 and it was shown that significant gains over utterance-based VTLN can be achieved with a small increase in computational complexity. Among the R-VTLN variants the algorithm using k-means for frame classification and region-independent warping factor computation was shown to be competitive in both performance and computational complexity. In the future we will investigate better criteria for selecting regions and how to combine R-VTLN with other normalization algorithms, e.g., bias removal.

7. ACKNOWLEDGMENTS

This work was partially funded by the EU FP6-IST project "HIWIRE". The authors wish to thank Prof. Richard Rose for many helpful discussions.

8. REFERENCES

[1] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," *Proc. ICASSP*, vol. 8, no. 2, pp. 353–356, May 1996.

[2] S. Panchapagesan and Ab. Alwan, "Multi-parameter frequency warping for vtln by gradient search," *Proc. ICASSP*, pp. 1181–1184, 2006.

[3] Ev. B. Gouvêa and R. M. Stern, "Speaker normalization through formant-based warping of the frequency scale," *Proc. Eurospeech '97, Rhodes, Greece*, vol. 8, no. 2, May 1997.

[4] P. Zhan and A. Waibel, "Vocal tract length normalization for large vocabulary continuous speech recognition," *Proc. ICASSP*, vol. 8, no. 2, May 1997.

[5] L. Welling, S. Kanthak, and H. Ney, "Improved methods for vocal tract normalization," *Proc. ICASSP*, pp. 761–764, March 1999.

[6] S. Molau, St. Kanthak, and H. Ney, "Efficient vocal tract normalization in automatic speech recognition," *Proc. EUROSPEECH*, pp. 2527–2530, March 1999.

[7] A. Miguel, E. Lieida, R. Rose, L. Buera, and A. Ortega, "Augmented state space acoustic decoding for modeling local variability in speech," *Proc. INTERSPEECH*, pp. 3009–3012, 2005.

[8] Ok Keun Shin, "A vector-quantizer based method of speaker normalization," *Proc. ICIS*, pp. 402–407, 2005.

[9] P. Zhan and M. Westphal, "Speaker normalization based on frequency warping," *Proc. ICASSP*, vol. 8, no. 2, pp. 1039–1042, 1997.