INTERSPEECH 2009 BRIGHTON

Acoustic Class Specific VTLN-Warping Using Regression Class Trees

S. P. Rath and S. Umesh

Department of Electrical Engineering,
Indian Institute of Technology Kanpur, Kanpur, India
[srath,sumesh]@iitk.ac.in

Abstract

In this paper, we study the use of different frequency warp-factors for different acoustic classes in a *computationally efficient* frame-work of Vocal Tract Length Normalization (VTLN). This is motivated by the fact that all acoustic classes do not exhibit similar spectral variations as a result of physiological differences in vocal tract, and therefore, the use of a single frequency-warp for the entire utterance may not be appropriate. We have recently proposed a VTLN method that implements VTLN-warping through a linear-transformation (LT) of the conventional MFCC features and efficiently estimates the warp-factor using the *same* sufficient statistics as that are used in CMLLR adaptation. In this paper we have shown that, in this framework of VTLN, and using the idea of regression class tree, we can obtain separate VTLN-warping for different acoustic classes. The use of regression class tree ensures that warp-factor is estimated for each class even when there is very little data available for that class. The acoustic classes, in general, can be any collection of the Gaussian components in the acoustic model. We have built acoustic classes by using data-driven approach and by using phonetic knowledge. Using WSJ database we have shown the recognition performance of the proposed acoustic class specific warp-factor both for the data driven and the phonetic knowledge based regression class tree definitions and compare it with the case of the single warp-factor.

Index Terms: VTLN, Acoustic-Class Specific Warping, Regression Class Tree, Linear Transform

1. Introduction

Inter-speaker variability is a major cause of performance degradation in speaker-independent (SI) speech recognition systems. Vocal Tract Length Normalization (VTLN) [1] is a commonly used method to reduce inter-speaker variability, where the spectra of the speech frames are appropriately frequency-warped to reduce the spectral variations among different speakers. It is a common practice to use a single warp-factor for the entire utterance that captures the global spectral variations among speakers. However, it is well known that all phone classes do not exhibit the same spectral variation due to physiological differences and hence it would be more appropriate to have different warp-factors for different acoustic classes within an utterance.

There has not been much work done on the use of class-specific warp-factors since conventional VTLN is cumbersome to implement. Even for the case of the single warp-factor, the estimation is performed by doing a grid search over a range of values of the warp-factor and selecting the one that maximizes the likelihood of the corresponding warped utterance w.r.t. the model. This approach is computationally expensive, since it requires generation of warped utterances for the entire search range of the warp-factor, which involves scaling of the filter-bank for each value of the warp-factor before generating the warped features.

One of the early works on the use of phone-specific warp-factors was reported in [2]. In this work, a preliminary phone transcription of the utterance obtained from first-pass recognition was used to assign a phoneme label to each acoustic vector and phoneme-dependent warping was estimated for each acoustic vector. However, no significant improvement was obtained over the use of a single warp-factor for the WSJ task.

Recently, the MATE [3] algorithm has been proposed where there is an expansion of the HMM state space to include the warp-factor space. In this method, frame-specific warp-factors are estimated. This allows the warp-factor to change every frame, but constraints are used to prevent abrupt changes in warp-factor for adjacent frames. Using this approach, they have shown 10% relative improvement over conventional VTLN on the Aurora2 task.

In MATE, Viterbi search is performed in a 3-dimensional space, that includes time, state, and warp-factor, i.e.,

$$\phi_{j,n}(t) = \max_{i \in I, \alpha_n \in A} \left\{ \phi_{i,m}(t-1) a_{i,j}^{m,n} \right\} b_j(x_t^{\alpha_n}) \quad (1)$$

where $\phi_{j,n}(t)$ is the likelihood of the optimum path terminating in state j and warp-factor α_n, I is the state-space and A is the warp-factor space. Essentially, in this method a frame-to-(state,warp-factor) mapping is obtained. Since the search is also performed along the warp-factor space, appropriately VTLN-warped acoustic vectors for *all warp-factors* are required to obtain the appropriate warp-factor for each frame. Other method which also finds multiple warp-factors from a speech utterance is [4].

Recently, we have shown that VTLN-warping can be implemented by a linear-transformation (LT) of the conventional MFCC features [5, 6]. These LT are analytically pre-computed without any modifications in the standard MFCC computations. Note that the linear transformation approaches proposed in [7, 8] are based on continuous-frequency domain processing and run into aliasing problems when implemented in practice. Further, our approach of LT VTLN exactly implements the conventional frequency-warping in the physical-frequency domain unlike the method proposed in [9], which, itself, is a modification of an earlier work of ours. Some of the other approaches to VTLN using linear transformation of MFCC are presented in [10, 11]. The use of these pre-computed matrices make VTLN-warping computationally efficient, since the VTLN-warped features can be obtained by a matrix-multiplication of MFCC features without the need of spectral-warping for each warp-factor, α, and generation of corresponding warped cepstral features.

In another earlier work [12], we have also shown that in the LT framework, it is also possible to estimate the warp factor very efficiently using the pre-computed VTLN matrices and the

same sufficient statistics as that are used in CMLLR. This makes VTLN even simpler.

In this paper, we show that in the computational efficient framework of [12], it is possible to estimate one warp-factor for each acoustic class, where the acoustic class, in general, can contain any collection of Gaussian components in the acoustic model. For example, if the acoustic class is defined as the collection of Gaussians belonging to a phoneme, a separate warp-factor can be obtained for each phoneme. The proposed method is similar to MATE since both the methods try to reduce the local variations in the speech spectra by using different warp-factors within an utterance. However, since the proposed method is based on acoustic class, it is flexible and computationally efficient. The salient features of the proposed method are:

1. A separate VTLN warp-factor is estimated for each acoustic class within an utterance.
2. The definition of acoustic class can be done in several ways, including using a distance criteria (data driven) or using phonetic knowledge.
3. A regression class tree can also be used to define regression classes. The advantage of using regression class tree is that when there is insufficient data available for accumulation of statistics for a particular acoustic class, its parent class can be used for warp-factor estimation of that class.
4. Linear-transformation approach of VTLN is used, and therefore, the generation of warped features by frequency-scaling is not necessary. Further, this approach is based on the idea of sufficient statistics similar to MLLR/CMLLR. This makes the warp-factor selection for each class very easy.
5. Since only one parameter needs to be estimated per acoustic class, very little adaptation data is required even when there are many acoustic classes unlike transform-based adaptation methods.

In Section 2 we briefly discuss our recently proposed LT approach for VTLN. In Section 3 we describe the sufficient statistics based approach for warp factor estimation, which will form the basis of the work presented in this paper. Then we explain our proposed method of regression class tree based warp-factor estimation in section 4. Finally, we present the experimental results on the WSJ database and compare word recognition performance of the proposed method for different number of regression classes.

2. Linear Transform Approach for VTLN

In [5], we have proposed a method to obtain VTLN-warped features, X^α, through a linear transformation of un-warped MFCC features, X, i.e.,

$$X^\alpha = W^\alpha X, \tag{2}$$

The VTLN warp-matrices, W^α, are obtained using the idea of band-limited interpolation and is given by

$$W^\alpha = DT^\alpha D^{-1} \tag{3}$$

where D is the DCT transform and T^α is the band-limited interpolation matrix given by

$$T^\alpha_{k,n} = \frac{1}{2N} \sum_{l=0}^{2N-1} e^{-j\frac{2\pi}{2N}\left(\frac{\tilde{\nu_l}}{\nu_s}\right)k} e^{j\frac{2\pi}{2N}\left(\frac{\nu_l}{\nu_s}\right)n}. \tag{4}$$

ν_l and $\tilde{\nu_l}$ denote the Mel-frequencies corresponding to the physical-frequencies (Hz) before and after frequency scaling, ν_s is the sampling frequency expressed in Mels and N is the number of Mel filters.

Since there is no closed form solution for the maximum likelihood estimation of the warp-factor, it is a common practice to perform a grid-search over the range of 0.8 to 1.2 to find the optimal α that maximizes the likelihood of the warped utterance w.r.t the model and the Jacobian term, i.e.,

$$\alpha^* = \arg\max_\alpha log\ p(W^\alpha X|\lambda, U) + log(|W^\alpha|) \tag{5}$$

where, U, λ and $|W^\alpha|$ are the transcription used for alignment, the model parameters and the Jacobian, respectively. Note that in this approach of warp-factor estimation, the likelihood has to be computed for each value of α. This requires multiple alignments of the utterance using Viterbi. Therefore, this approach of warp-factor estimation is computationally expensive.

3. Sufficient Statistics based Method for Global Warp-Factor Selection

Recently [12], we have shown that in the linear transformation framework, the warp factor selection can be done very efficiently using two sufficient statistics that are collected over the speech utterance and the pre-computed VTLN matrices described in Section 2. Below we summarize the steps involved to estimate the warp-factor of an utterance.

Initial Step: Compute and store the VTLN warp-matrices using Eq. 3 and 4. (Note: These matrices are data independent.)

1. Obtain the posterior probability, $\gamma_{jm}(t)$, of the un-warped features w.r.t. the model.
2. Compute the following two statistics over all Gaussian components in the acoustic model using the un-warped feature vectors, i.e.,

$$K^{(i)} = \sum_{m=1}^{M} \frac{\mu^{(i)}_{jm}}{\sigma^{(i)^2}_{jm}} \sum_{t=1}^{T} \gamma_{jm}(t)\, x_t{}^T \tag{6}$$

$$G^{(i)} = \sum_{m=1}^{M} \frac{1}{\sigma^{(i)^2}_{jm}} \sum_{t=1}^{T} \gamma_{jm}(t)\, x_t x_t{}^T. \tag{7}$$

   $\mu^{(i)}_{jm}$, $\sigma^{(i)\,2}_{jm}$ and M are the mean, the co-variance and total number of Gaussian components in the model, respectively.
3. To get the α for the utterance, perform a simple maximization over the warp-matrices, i.e.,

$$\alpha^* = \arg\max_\alpha J - \frac{1}{2}\left\{\sum_{i=1}^{D} w_i^\alpha G^{(i)} w_i^{\alpha^T} - 2K^{(i)} w_i^{\alpha^T}\right\} \tag{8}$$

   where, J, w_i^α and D are the Jacobian, the i^{th} row of W^α and the dimension of feature vectors, respectively.

The statistics, K and G, are exactly the *same sufficient statistics* as that are used in CMLLR. Therefore, the method is very similar to CMLLR and can be used as conveniently as any transform-based speaker adaptation methods. However, in VTLN, only *one* of the pre-computed matrices, corresponding to the warp-factors, is selected. On the other hand, in the case of CMLLR, the elements of the adaptation matrix (usually block-diagonal) are estimated. Therefore, VTLN requires very little data to achieve speaker normalization.

4. Regression Class Tree based VTLN

Since the statistics used in Eq. 6 and 7 are collected over all Gaussian components in the acoustic model, the transform obtained is a *global* transform, which is applied on all components in the acoustic model. However, it is also possible to cluster similar components into an acoustic class and accumulate the statistics, G and K, separately for the class. In this scenario, therefore, it is possible to obtain a separate warp-factor for each class. For warp-factor estimation, the steps given in Section 3 for the case of global warp-factor remain the same, except that the following two modified statistics are used:

$$K_r^{(i)} = \sum_{m_r=1}^{M_r} \frac{\mu_{jm_r}^{(i)}}{\sigma_{jm_r}^{(i)^2}} \sum_{t=1}^{T} \gamma_{jm_r}(t)\, x_t{}^T \tag{9}$$

$$G_r^{(i)} = \sum_{m_r=1}^{M_r} \frac{1}{\sigma_{jm_r}^{(i)^2}} \sum_{t=1}^{T} \gamma_{jm_r}(t)\, x_t x_t{}^T \tag{10}$$

where, the outer summation is performed over all Gaussian components belonging to r^{th} acoustic class.

Since only one parameter is required to be estimated for each class, even with limited data, we can have multiple classes. In this paper, we have used the ideas of regression classes and regression class trees [13] for warp-factor estimation. The advantage of using regression class tree is that even in the case of insufficient data available for a class, warp-factor can still be estimated using its parent regression class. In our experiments, the following two methods are used to define the regression trees.

4.1. Phonetic Knowledge Based Approach

In this approach, a set of broad phonetic classes is defined and Gaussian components belonging to each broad phonetic class are clustered together to form a regression class. Table 1 shows the broad phonetic classes (excluding silence class) that are used in this paper for the experiments. This phoneme list is similar to that is used in [14].

Table 1: *Broad phonetic classes used in the experiments [14].*

Broad Phonetic Classes	Phonemes
Very Front Vowel	ih,iy
Near Front Vowel	ae, eh
Front Diphthongs	ey, ay
Back Diphthongs	aw, ow, oy
Near Back Vowel	aa, uh, ah, er
Very Back Vowel	ao, uw
Liquids	l, r, w, y
Nasals	m, n, ng
Strong Fricatives	dh, jh, v, z, zh
Weak Fricatives	ch, f, hh, s, sh, th
Unvoiced Stops	k, p, t
Voiced Stops	b, d, g

4.2. Data Driven Based Approach

In this approach, a binary regression tree was constructed by successively splitting Gaussian components clustered at a node in the tree. To start with, all components in the model are clustered at the global node. Using the centroid splitting algorithm, two child nodes are formed from a node and using the euclidean distance as the measure of similarity, each Gaussian component in the node is assigned to one of the child nodes. This procedure is continued until the required number of classes are formed. The tree is built using the SI/previous-iteration VTLN model.

5. Experimental Set up

We present the experimental results on the Wall Street Journal (WSJ) database. Cross word tri-phone models were used with decision tree based state tying. The tri-phone models had three states, with 8 diagonal-covariance components for each state. A three state model with 16 diagonal-covariance components was used for the silence, and a short-pause model (allowing skip) was constructed with all states tied to the silence model. Each component was modeled by a Gaussian density function. The acoustic models were trained using the WSJ0-84 training set that resulted in 2736 states after doing state tying. Test was performed on November 1992 WSJ test set using WSJ 5K closed non-verbalized vocabulary and the WSJ 5K closed non-verbalized bi-gram language model. The features in the task are 39-dimensional MFCC, comprising normalized log-energy, $c_1, \ldots, c_{12}$ (excluding c_0) and their first and second order derivatives. 20 ms frames with 10 ms overlap was used and cepstral mean subtraction was applied over every speech utterance. All experiments were conducted using Hidden Markov Models Toolkit (HTK).

6. Results and Discussions

Now we present the experimental results for our proposed method of regression class tree based warp-factor estimation. In Table 2 the Word Recognition Accuracy (WRA) for the distance based regression class trees are shown. Global Class in the table indicates the case where one global warp-factor was estimated for the utterance. In this case silence was also warped by the global warp-factor along with the speech components. However, we found that it is better not to warp silence in the other cases. The numbers in the first column of the table show the number of regression classes that were used for warp-factor estimation for each utterance, excluding the silence class. For example, in the case of 8 classes in the table, silence was excluded and a regression class tree was grown with 8 speech classes using the data driven approach, and warp-factor estimation was performed. The minimum frame count for warp-factor estimation was kept to 30 frames. The column "Adaptation" indicates that the un-normalized model was used for warp-factor estimation and in the subsequent recognition of test data. "Adaptive Training" indicate that VTLN was also done during training and the speaker-normalized model was used during test.

Table 2: *Word Recognition Accuracy for different number of regression classes in data driven approach.*

Number of Regression Classes (excluding silence)	Adaptation	Adaptive Training
Global Class	93.75	94.00
2	93.83	94.02
4	93.83	94.10
6	93.81	94.04
8	93.85	94.10
10	93.88	94.25
12	93.90	94.23

• No Normalization (Baseline) performance: 93.60

From Table 2, the following observations can be made.

• In the case of multiple regression classes, the WRA performance is higher than that of the global class.

• There is a trend for WRA to increase with increasing number of acoustic classes. The highest WRA was obtained for the case of 10 classes.

Table 4: *Warp-factor distribution for the broad phonetic classes used in the experiments and the corresponding global warp-factor shown for a few example utterances taken from the training set of WSJ0-84.*

	Global Class	Very Front Vowel	Near Front Vowel	Front Diphthongs	Back Diphthongs	Near Back Vowel	Very Back Vowel	Liquids	Nasals	Strong Fricatives	Weak Fricatives	Unvoiced Stops	Voiced Stops
example 1	1.04	1.02	1.02	1.04	1.04	1.04	1.04	1.06	1.02	1.06	1.08	1.10	1.06
example 2	1.06	1.04	1.06	1.12	1.06	1.08	1.06	1.04	1.08	1.10	1.06	1.10	1.10
example 3	0.94	0.96	0.96	0.96	0.96	0.94	0.94	0.96	0.94	0.90	0.92	0.92	0.90
example 4	0.94	0.96	0.96	0.94	0.94	0.94	0.94	0.92	1.06	0.92	0.92	0.92	0.94
example 5	0.98	0.98	1.00	0.98	0.98	0.98	0.98	1.00	1.02	1.02	0.98	0.96	0.98

Table 3: *Word Recognition Accuracy for phonetic knowledge based approach of acoustic class.*

Number of Regression Classes (excluding silence)	Adaptation	Adaptive Training
12	93.87	94.17

Table 3 shows the experimental results conducted using the phonetic knowledge based approach for regression class definition. The broad phonetic classes shown in in Table 1 were used for the experiments. Here also, silence was considered as a separate class and was not warped. The minimum frame counts for warp-factor estimation was kept to 30 frames. We observe that in this approach also, the performance using 12 acoustic class is better than the global class case (shown in Table 2). Comparing Table 2 and Table 3, we observe that the WRA of the phonetic knowledge based approach is comparable to the data driven approach for the case of 12 classes.

6.1. Analysis of Warp factors for different classes

In Table 4 the warp-factors for each of the broad phonetic class are shown for some of the utterances taken from the training set of the WSJ database. The following observations can be made:

• There are variations in the warp-factor among different phoneme classes in the speech utterance validating our conjecture that all phone classes do not undergo similar spectral variations.

• Similar phoneme classes (for example, Very Front Vowel and Near Front Vowel or Very Back Vowel and Near Back Vowel) have similar warp-factors, even though they were estimated using different regression classes.

• Unvoiced Stops, Strong Fricatives etc. have significantly different warp-factors than the global warp-factor.

7. Conclusions

In this paper we have proposed a method for using regression class trees in VTLN to allow acoustic class specific warp-factor estimation. The proposed method is very flexible since the acoustic class can be chosen in different ways and supports the use of regression class trees. From experiments performed on the WSJ database, it was observed that using multiple acoustic classes for warp-factor estimation provided better recognition performance than using one global acoustic class. This observation holds true both in data driven and phonetic approaches of regression class tree definitions. Our experiments also indicate that the warp-factors are different for different acoustic classes. In all our experiments, VTLN is efficiently implemented using linear-transformation of MFCC and using the same sufficient statistics as CMLLR for warp factor estimation. It is, therefore, computationally more efficient than MATE or the method of applying phone-specific warping using alignment output.

8. Acknowledgments

A part of this work was supported by SERC project funding `SR/S3/EECE/0058/2008` from the Department of Science & Technology, Ministry of Science & Technology, India.

9. References

[1] L. Lee and R. Rose, "Frequency Warping Approach to Speaker Normalization," *IEEE Trans. on Speech and Audio Processing*, vol. 6, pp. 49–59, January 1998.

[2] S. Molau, S. Kanthak, and H. Ney, "Efficient Vocal Tract Normalization in ASR," in *Proc.ESSV*, Cottbus, Germany, 2000.

[3] A. Miguel, E. Lleida, R., R. L. Buera, and A. Ortega, "Augmented State Space Acoustic Decoding for Modeling Local Variability in Speech," in *Interspeech05*, Lisbon, Portugal, 2005.

[4] M. G. Maragakis and A. Potamianos, "Region-Based Vocal Tract Length Normalization for ASR," in *Interspeech*, Brisbane, 2008.

[5] D. R. Sanand and S. Umesh, "Study of Jacobian Compensation Using Linear Transformation of Conventional MFCC for VTLN," in *Interspeech2008*, Brisbane, Australia, 2008.

[6] S. Umesh, A. Zolnay, and H. Ney, "Implementing Frequency Warping and VTLN Through Linear Transformation of Conventional MFCC," in *Interspeech2005*, Lisbon, Portugal, 2005.

[7] M. Pitz and H. Ney, "Vocal Tract Normalization Equals Linear Transformation in Cepstral Space," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 930–944, 2005.

[8] J. McDonough, W. Bryne, and X. Luo, "Speaker Normalization with All-Pass Transforms," in *Interspeech98*, Sydney, 1998.

[9] S. Panchapagesan, "Frequency Warping by Linear Transformation of Standard MFCC," in *Interspeech*, Pittsburgh, USA, 2006.

[10] T.Claes, I.Dologlou, L. Bosch, and D. van Compernolle, "A Novel Feature Transformation for Vocal Tract Length Normalisation in ASR," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 6, pp. 549–557, 1998.

[11] S. Cox, "Speaker Normalization in the MFCC Domain," in *Interspeech00*, Beijing, China, 2000.

[12] P. T. Akhil, S. P. Rath, S. Umesh, and D. R. sanand, "A Computationally Efficient Approach to Warp Factor Estimation in VTLN using EM Algorithm and Sufficient Statistics," in *Interspeech2008*, Brisbane, Australia, 2008.

[13] M. J. F. Gales, "The Generation and Use of Regresion Class Trees for MLLR Adaptation," in *Tech. Rep. CUED/F-INFENG/TR263*, Cambridge Univ., Cambridge, U. K., 1996.

[14] C. J. Leggetter, "Improved Acoustic Modelling for HMMs Using Linear Transformations," Ph.D. dissertation, University of Cambridge, UK, 1995.