Using VTLN Matrices for Rapid and Computationally-Efficient Speaker Adaptation with Robustness to First-Pass Transcription Errors

S. P. Rath, S. Umesh and A. K. Sarkar

Department of Electrical Engineering, Indian Institute of Technology Kanpur, Kanpur, India

Abstract

In this paper, we propose to combine the rapid adaptation capability of conventional Vocal Tract Length Normalization (VTLN) with the computational efficiency of transform-based adaptation such as MLLR or CMLLR. VTLN requires the estimation of only one parameter and is, therefore, most suited for the cases where there is little adaptation data (i.e., rapid adaptation). In contrast, transform-based adaptation methods require the estimation of matrices. However, the drawback of conventional VTLN is that it is computationally expensive since it requires multiple spectral-warping to generate VTLN-warped features. We have recently shown that VTLN-warping can be implemented by a linear-transformation (LT) of the conventional MFCC features. These LTs are analytically pre-computed and stored. In this frame-work of LT VTLN, computational complexity of VTLN is similar to transform-based adaptation since warp-factor estimation can be done using the same sufficient statistics as that are used in CMLLR. We show that VTLN provides significant improvement in performance when there is small adaptation data as compared to transform-based adaptation methods. We also show that the use of an additional decorrelating transform, MLTT, along with the VTLN-matrices, gives performance that is better than MLLR and comparable to SAT with MLTT even for large adaptation data. Further we show that in the mismatched train and test case (i.e., poor first-pass transcription), VTLN provides significant improvement over the transform-based adaptation methods. We compare the performances of different methods on the WSJ, the RM and the TIDIGITS databases.

Index Terms: VTLN, Rapid Adaptation, MLTT, CAT, Linear Transform

1. Introduction

In recent years, there has been a lot of interest in reducing inter-speaker variability to improve the performance of Speaker-Independent (SI) speech recognition systems. A popular approach to reduce inter-speaker variability is to use transform-based adaptation such as MLLR and CMLLR [1, 2], where the parameters of the model are adapted to a particular speaker. These methods require the estimation of the elements of the matrix from the adaptation data and is usually done only during the test phase. Speaker Adaptive Training (SAT) [2] is a slightly different form of adaptation in that a speaker normalized model is formed by estimating CMLLR matrices during training also. Because of the ease of applicability, these transform-based adaptation schemes are simple and have been widely used in many speech recognition systems. However, even in the case of block-diagonal matrices (one block each for static, \(\Delta\), and \(\Delta \Delta\) coefficients), sufficiently large amount of adaptation data is required for reliable estimation of the elements of the adaptation matrix. Often, even without regression classes, two transforms are used - one for speech and one for silence. Hence there has been a lot of interest in reducing the number of parameters to be estimated during adaptation by using the idea of Eigen-voice [3], Cluster Adaptive Training (CAT) [4] etc. This is particularly important in applications where the user may provide only a small amount of speech data.

VTLN [5] is another commonly used method to reduce inter-speaker variability. VTLN is achieved by appropriately frequency-warping the spectra before computing the cepstral features. The frequency warp-factor for an utterance is found by maximizing the likelihood of the warped utterance using the model and first-pass transcription, i.e.,

\[
\alpha^* = \arg \max_{\alpha} \log p(X^\alpha|\lambda, U)
\]

where, \(X^\alpha, U\) and \(\lambda\) are the frequency-warped cepstral features, the transcription, and the model parameters, respectively. Since VTLN requires the estimation of only one parameter, namely \(\alpha\), it requires very little adaptation data. However, the implementation of conventional VTLN is cumbersome since it requires computation of the warped features, \(X^\alpha\), for a range of values (0.8 to 1.2) of \(\alpha\) after appropriate frequency-warping. This involves scaling of the filter-bank to generate warped features.

Recently, we have shown that VTLN-warping can be implemented by a linear transformation (LT) of the conventional MFCC features. These LTs are analytically pre-computed and stored [6, 7]. Unlike [8, 9], which are derived in continuous frequency-domain and hence suffer from aliasing, our proposed LT is obtained directly in the discrete-domain. Further, our method does exact conventional frequency-warping unlike the method of [10], which is a modification of an earlier work of ours. Other method which approximates the LTs includes [11]. The use of these pre-computed matrices make VTLN attractive again since we need to find only the optimal pre-computed warp matrix during \(\alpha\)-estimation followed by the application of this LT to get normalized features.

Recently, we have also proposed a method where using the same sufficient statistics that are used in CMLLR and these pre-computed VTLN matrices, it is also possible to estimate the warp factor very efficiently [12]. In this framework, VTLN is even simpler.

VTLN, in its LT form, resembles closely SAT [2] adaptation, since both are feature domain transformations and in both methods, speaker-normalized models are used in the recognition phase. In this paper, we show that by using VTLN (or SAT) matrices along with the MLTT [13] decorrelating transform, we can get improvement in the performance of VTLN.
We have found that, while the performance of SAT without MLLT is usually superior compared to VTLN without MLLT, the gain in performance of VTLN with MLLT is significantly higher, making the performance of VTLN with MLLT comparable to SAT with MLLT. More importantly, the performance of VTLN with MLLT is much better than that of MLLR and CAT, even for the case when there is sufficient adaptation data to estimate MLLR matrices. The use of MLLT does not add any additional computation during the test phase as the matrix is completely estimated from the training data during the training phase. Finally, we show that when there is very little adaptation data (e.g. a single utterance) and adaptation transforms cannot be estimated, VTLN with MLLT still works well and gives good improvement in performance.

In this paper, we also show that in the mismatched train and test conditions, VTLN with MLLT significantly outperforms all transform-based adaptation methods, even when there is sufficient adaptation data. This is because transform-based adaptation methods are sensitive to first-pass transcription errors.

In Section 2 we discuss our recently proposed LT approach for VTLN followed by a brief description of various transform-based speaker adaptation methods that we use in the paper. In Section 4 we give an overview of MLLT. Then, we present our experimental results on different databases and compare the word recognition accuracy performance using different methods and under different cases of available adaptation data.

2. Linear Transform Approach to VTLN

VTLN is a feature-space transformation that modifies the spectra of the speech signal by frequency-warping to reduce interspeaker variability [5]. Usually, a grid-search is performed to select the warp-factor for each utterance, which requires the generation of warped features for a range of values of warp-factor. We have recently shown that warped features, $X^a$, can be generated using a LT of the un-warped MFCC features [6], $X$, i.e.,

$$X^a = W^a X = D T^a D^{-1} X$$

where the warp matrix, $W^a$, can be analytically computed using $D$, the DCT transform, and $T^a$, the band-limited interpolating matrix, which is given by

$$T^a_{k,n} = \frac{1}{2N} \sum_{l=0}^{2N-1} e^{-j 2\pi (\frac{\nu}{N}) k} e^{j 2\pi (\frac{\nu}{N}) n}.$$  

$\nu_l$, $\nu_1$, $\nu_2$, and $N$ denote the Mel-frequencies corresponding to the Hz-frequencies after and before frequency scaling, the sampling frequency in Mel-scale and the number of Mel filters, respectively. Hence it is only necessary to generate the un-warped ($a=1.0$) MFCC features. The warped features for other values of $a$ can be generated using Eq. 2. In this paper, the linear-transformations of VTLN, as described by Eq. 2 and 3, is used in the experiments. In this framework of LT VTLN, the warp-factor can be efficiently estimated as described next.

2.1. An Efficient Method for Warp-Factor selection

We have recently shown that [12] in LT based VTLN it is possible to obtain the warp-factor using exactly the same sufficient statistics as that are used in CMLLR, i.e.,

$$\alpha^* = \arg \max_{\alpha} J = \frac{1}{2} \sum_{i=1}^{N} w_{i}^a G^{(i)} w_i^a T - 2K^{(i)} w_i^a T$$

where $w_{i}^a$ and $J$ are the the $i^{th}$ row of the $W^a$ matrix and the Jacobian, respectively, and

$$K^{(i)} = \sum_{m=1}^{M} \sum_{j=1}^{T} \gamma_m (t) x_i x_T$$

$$G^{(i)} = \sum_{m=1}^{M} \sum_{j=1}^{T} \gamma_m (t) x_i x_T$$

($K^{(i)}$ and $G^{(i)}$ are same as CMLLR), $\mu_{jm}^{(i)}$, $\sigma_{jm}^{(i)}$, $\gamma_m (t)$ and $M$ are the mean, the co-variance, the posterior probability of the un-warped features and the total number of components in the acoustic model, respectively. In this frame-work, VTLN is as simple to use as any other transform-based speaker adaptation methods. This method will henceforth be denoted as EMMSS (Expectation-Maximization Sufficient-Statistics). The like-lihood can also be computed by Viterbi alignment, in which case it reduces to conventional VTLN without separate re-alignments for each $\alpha$. This Viterbi based VTLN will be referred to as Viterbi VTLN.

3. Review of Speaker Adaptation Methods

We briefly, describe several Speaker Adaptation Methods that we have used in this paper. In Maximum Likelihood Linear Regression (MLLR) [1], only the means of the SI model are transformed through an affine transformation, $\mu = A \mu + b$, to match the test speaker’s speech. And in constrained MLLR (CMLLR) [2], the same matrix is used to transform both the mean and the covariance, i.e., $\mu = A \mu + \Sigma = A \Sigma A^T$. Adaptation data is used to estimate the MLLR/CMLLR transforms $(A, b)$ in ML sense using the first-pass transcription and the SI model. When CMLLR adaptation is used both in training and test, it is referred to as SAT [2]. It is a common practice to have a separate adaptation transform for the speech part and another transform for the silence part. In this paper, we have conducted experiments considering different matrix structures for $A$, and with and without the use of bias, $b$.

In CAT [4], speaker clustered models are trained from the training speakers. Using the adaptation data, a linear combination of means of these models is obtained in the ML sense. CAT is a rapid speaker adaptation method since only few weighting coefficients have to be estimated from the adaptation data.

4. Maximum Likelihood LT (MLLT)

MLLT [13] is a special case of Heteroscedastic Linear Discriminant Analysis (HLDA) [14], which is a class discriminant analysis method that does not assume equal co-variance between the classes. When HLDA is configured to have no reduction of feature dimension, it works like a decorrelating transform for the feature vectors and is known as MLLT. We have used MLLT matrices in our experiments both with VTLN and SAT. In the case of VTLN, a global MLLT is estimated from the training data after they are VTLN-warped (or, in case of SAT, transformed by the CMLLR matrices) by the warp-factors of the respective speakers in training set. MLLT transformation helps in decorrelating the feature vectors as well as adapting to any covariance mismatch between transformed features and the model. MLLT estimation is entirely done during training stage and no MLLT estimation is necessary during test in any of our experiments. In our experiments, we will show that the use of VTLN along with MLLT gives performance similar to the transform-based methods.
5. Results and Discussions

We present the experimental results on the Wall Street Journal (WSJ), the DARPA Resource Management (RM) and the TIDIGITS databases. In the case of WSJ, cross-word tri-phone models were used with decision tree based state tying. The tri-phone HMM models consist of three states, with 8 diagonal-covariance (Gaussian) components per state. A three state model with 16 diagonal-covariance components was used for silence ("sil"), and the short-pause ("sp") model (allowing skip) was constructed with all states tied to silence model. The acoustic models were trained using the WSJ0-84 training set that resulted in 2736 states after doing state tying. Test was done on Nov-92 WSJ test set using WSJ 5K closed non-verbalized dictionary and the WSJ 5K closed bi-gram language model. In the RM, the acoustic models were trained similarly as the WSJ except that 6 mixtures were used per state and the single-state "sp" model was tied to the middle state of the "sil" model. Training was done using the RM SI-109 training set that resulted in 1560 states after state tying. Test was performed on the Feb-89 test set using RM word pair language model. In case of the TIDIGITS, 11 digits models with 16 states and 5 diagonal covariance components were used. "sil" and "sp" models were similar to RM. All adaptations were performed in unsupervised mode. The features in all tasks are 39-dimensional MFC, comprising normalized log-energy, $c_1, \ldots, c_{12}$ and their first and second order derivatives. 20 ms frames with 10 ms overlap was used and cepstral mean subtraction was applied on every utterance. All experiments were conducted using HTK.

### 5.1. Experiments on RM Task

The Word Recognition Accuracy (WRA) results of different adaptation experiments on the RM task database are presented in Table 1. Results are broken up into two broad classes, namely, Adaptation and Adaptive Training. While in Adaptation the un-normalized model is used, in Adaptive Training the adaptation of training speakers are done to obtain a speaker-normalized model. In the cases of Adaptation, Normalized models are trained using all training utterances from a particular training-speaker to obtain the adaptation transform (warp-factors in case of VTLN) for that speaker. During the test phase of both Adaptation and Adaptive Training, the number of utterances for adaptation were varied to study the effect of the amount of adaptation data on the different methods. The first column of Table 1 shows the amount of adaptation data used for transform/warp-factor estimation during test. The average number of frames per test utterance was $\approx 330$ frames ($3.3s$). To ensure robust estimation of MLLR/CMLLR matrices, the minimum frame count was kept to 1000 for block-diagonal case and 350 for diagonal case. However, for the single utterance case, the minimum frame count is not met for the block-diagonal case and the performance remains same as that of the baseline. In the case of VTLN we show the performance using Viterbi and EM-SS methods described in Section 2.1. Table 1 also shows the performance with and without the use of MLLT. Finally, for the CAT experiments, a model based approach was followed with 2 clusters (no bias cluster). We make the following observations based on the WRA performance of different methods.

- The performances of VTLN using Viterbi and EM-SS are comparable. VTLN without MLLT (both for Viterbi and EM-SS in Adaptive Training) is better than MLLR-Diag and CMLLR-Diag but slightly inferior compared to Blk-Diag case of MLLR, CMLLR and SAT. VTLN in adaptation-only framework is comparable to the Diag-cases of MLLR and CMLLR.
- However, performance of VTLN with MLLT is comparable to SAT with MLLT and better than MLLR and CAT for all cases of adaptation data.
- More importantly VTLN with MLLT based adaptation gives significant gain in adaptation performance for small amount of adaptation data (e.g. 1 or 2 utterances) when transform-based adaptation give little or no gain. This is because VTLN needs only the warp parameter to be estimated from the adaptation data. In these cases, performance of CAT is better than MLLR/CMLLR, but less than VTLN with MLLT.

### 5.2. Experiments on WSJ Task

Similar experiments were conducted on the WSJ database and results are shown in Table 2. We found that the VTLN results using EM-SS are slightly inferior to Viterbi method. Hence only VTLN-Viterbi results are shown. Similarly, performance of SAT was consistently better than CMLLR and hence only SAT results are shown. The average number of frames/utterance in the test set was $\approx 734$ frames ($7.34s$) and minimum frame count was kept to 1000 for MLLR/SAT cases. The followings are observed from the table.

- In the cases of use of large adaptation data, the performance of VTLN in adaptive training framework with MLLT is better than MLLR, and comparable to SAT with MLLT.
- In the case of small adaptation data (e.g. 1 utterance), WRA of VTLN is significantly better than all other methods.

### 5.3. Adaptation for Mismatched Train and Test Speakers

We have conducted a set of experiments on TIDIGITS database where there is a mismatch between train and test speaker con-
Table 2: Word Recognition Accuracy for WSJ Task.

<table>
<thead>
<tr>
<th># of utt. used in test</th>
<th>No MLLT</th>
<th>With MLLT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Adapt.</td>
<td>Adapt. Training</td>
</tr>
<tr>
<td></td>
<td>MLLR</td>
<td>SA I</td>
</tr>
<tr>
<td></td>
<td>Blk B</td>
<td>Blk NB</td>
</tr>
<tr>
<td>40</td>
<td>94.50</td>
<td>94.92</td>
</tr>
<tr>
<td>20</td>
<td>94.46</td>
<td>94.85</td>
</tr>
<tr>
<td>10</td>
<td>94.37</td>
<td>94.85</td>
</tr>
<tr>
<td>8</td>
<td>94.23</td>
<td>94.94</td>
</tr>
<tr>
<td>5</td>
<td>94.25</td>
<td>94.73</td>
</tr>
<tr>
<td>4</td>
<td>94.35</td>
<td>94.67</td>
</tr>
<tr>
<td>2</td>
<td>94.12</td>
<td>94.56</td>
</tr>
<tr>
<td>1</td>
<td>93.94</td>
<td>94.17</td>
</tr>
</tbody>
</table>

Baseline Performance=93.60, Adapt.=Adaptation

Table 3: Word Recognition Accuracy for TIDIGITS Task.

Column 1 shows train and test speaker conditions. For e.g., in the case of Male-Child (M-C), male speakers were used for training and data from children were used in test. Similar experiments were done with Male-Female. The average number of frames/utterance in the test set was ≈ 200 frames (2.0s). In these cases of mismatched train-test speaker conditions, for example in M-C, the MLLT transform was estimated using warped features/CMLLR matrices of the children train data. From Table 3 following observations can be made:

- In the cases of M-C, the performance of SAT and MLLR are significantly worse than VTLN without MLLT. After the use of MLLT, there is a significant improvement in performance of SAT with MLLT. However, VTLN with MLLT still gives better WRA than SAT with MLLT.
- The poor performance of SAT and MLLR when compared to VTLN for the M-C case is due to the poor first-pass transcription (only 68.39%). With the use of true transcription, MLLR and CMLLR have performances 99.14% and 99.10%, respectively for the case of 7 utterances/speaker. Thus, while VTLN is insensitive to transcription errors, CMLLR/CMLLR matrix estimation is very sensitive to such errors.
- In the case of M-F, MLLR and SAT have comparable performance to VTLN without MLLT when enough adaptation data (77, 11 utterances) is available but perform poorly for low adaptation data (1 & 7 utterances) due to insufficient data for matrix estimation with minimum frame count of 1000. Note that in this case the first-pass transcription is good. After MLLT, however, SAT and VTLN performed similarly in all cases.
- For single utterance cases, VTLN gave significant improvement in performance. MLLR and SAT matrices are not updated due to shortage of adaptation data.

6. Conclusions

In this paper we have proposed to use LT based VTLN matrices for adaptation. The primary motivation being that very little adaptation data is required for estimating the single warp-parameter. It also has the advantage of computational efficiency unlike conventional VTLN since the warp-factor estimation can be done using the same statistics as used in CMLLR. We have compared the WRA performance of our method with MLLR and SAT by conducting experiments on WSJ, RM and TIDIGITS databases. We showed that with the use of MLLT our method performed better than MLLR and was similar to SAT with MLLT when amount of adaptation data was large. Further, in the mismatched train-test conditions MLLR/SAT perform significantly worse than our proposed approach due to poor first-pass transcription. Finally, when the amount of adaptation data is small, SAT or MLLR cannot be used because the matrix cannot be estimated and yet our proposed VTLN-based adaptation provides significant improvement.

7. Acknowledgments

A part of this work was supported by SERC project funding SR/S3/EECE/058/2008 from the Department of Science & Technology, Ministry of Science & Technology, India.

8. References