A Study on Combining VTLN and SAT to Improve the Performance of Automatic Speech Recognition

D. R. Sanand and Mikko Kurimo

Adaptive Informatics Research Center, Aalto University, Finland.
[rama.doddipatla,mikko.kurimo]@tkk.fi

Abstract

In this paper, we present ideas to combine VTLN and SAT to improve the performance of automatic speech recognition. We show that VTLN matrices can be used as SAT transformation matrices in recognition, though the training still follows conventional SAT. This will be useful when there is very little adaptation data and the SAT transformation matrix can not be estimated to perform the required adaptation. We also present a study to understand whether VTLN can be performed after SAT and whether such a combination is better than the conventional approach, where VTLN is performed before SAT. Finally, we present a novel approach to perform VTLN by using VTLN matrices in cascade. This allows us to include warping-factors that are not included in the initial search space. We show through recognition experiments that these combinations improve the performance of ASR, with major gains in the mis-matched train and test speaker conditions.

Index Terms: VTLN, SAT, Linear Transformation, Automatic Speech Recognition

1. Introduction

Inter-speaker variability is a major source of performance degradation in speaker independent (SI) automatic speech recognition (ASR). The techniques to handle speaker variability have been broadly classified as speaker normalization and speaker adaptation in ASR literature. Speaker normalization modifies the front-end, whereas speaker adaptation modifies the model parameters to normalize speaker variability and there by reducing the mis-match between the trained model and the test speaker.

Vocal tract length normalization (VTLN) [1, 2] is a widely used speaker normalization approach, that minimizes the variations in the speech spectra arising due to the differences in the vocal tract lengths (VTL's) of speakers uttering the same sound. Speaker variability is accounted by performing spectral scaling. Speaker adaptation approaches on the other hand, estimate a transformation matrix from the data to transform the model parameters (means and covariances) to account for speaker variability. Some of the widely used adaptation approaches include: maximum likelihood linear regression (MLLR) [3] and constrained MLLR (CMLLR) [4, 5]. CMLLR transforms both means and covariances using a single transformation and MLLR only the means or the means and covariances using separate transformations [6].

In this paper, our interest is primarily focused on VTLN and CMLLR and how they can be combined to improve the performance of SI-ASR. The reason for choosing these transformations is that they both can be expressed as feature transformations. CMLLR can be seen as a feature transformation matrix due to the constraint that the same matrix is used for transforming both means and covariances, i.e.

$$\mathcal{L}(\mathbf{X};\mu,\mathbf{\Sigma},\mathbf{B}) = \mathcal{L}(\mathbf{B^{-1}X};\mu,\mathbf{\Sigma}) + log(|\mathbf{B^{-1}}|) \quad (1)$$

The above equation states that, the likelihood of the feature $\mathbf{X}$ given the model parameters mean (μ) and covariance (Σ) along with the transformation ($\mathbf{B}$) is equivalent to transforming the features with the inverse transformation and accounting for Jacobian [5].

Recently, it has been shown that VTLN can be represented as a linear transformation on conventional mel frequency cepstral coefficients (MFCC) [7, 8], i.e.

$$\mathbf{X}^{\alpha} = \mathbf{A}^{\alpha}.\mathbf{X} \quad (2)$$

where, $\mathbf{X}^{\alpha}$ represent the VTLN warped MFCC features and $\mathbf{A}^{\alpha}$ represent the VTLN transformation matrix for a specific warping factor α. $\mathbf{A}^{\alpha}$ can be treated as a CMLLR transformation matrix that is pre-computed rather than estimated form the data and performs only spectral scaling.

VTLN is usually performed in both training and recognition. Similarly, CMLLR can also be performed in both training and recognition, which is know in literature as speaker adaptive training (SAT) [9]. It is easy to understand that, if sufficient data has been presented to estimate the transformation matrix robustly, CMLLR usually obtains a higher performance than VTLN. This is because, CMLLR has more free parameters to describe the characteristics of the data.

In this paper, we will present our investigations to combine VTLN with SAT and how these combinations influence the performance of SI-ASR. In this direction, we present an idea to use VTLN matrices as SAT transformation matrices in recognition. The need for such a combination arises when there is very little adaptation data and a SAT transformation can not be estimated robustly. Next, we investigate the idea of performing VTLN on top of SAT. It will be interesting to understand whether such a combination is advantageous over the conventional approach, where VTLN is performed before SAT. We also present a novel approach to VTLN by using the VTLN matrices in cascade. This allows us to include warping-factors that are not seen in the initial search space.

The rest of the paper is organized as follows: First, we present a brief overview of VTLN and SAT. Then, present our investigations on combining VTLN with SAT and how do they affect the recognition performance on both matched and mismatched train and test speaker conditions. All our experiments are based on Wall Street Journal (WSJ0) task. Finally, we present our conclusion.

2. Vocal Tract Length Normalization

VTLN is a simple approach to speaker normalization and requires the estimation of a single parameter (α) that controls the amount of spectral scaling. In order to estimate this single parameter a maximum likelihood (ML) based grid search is performed over a pre-defined range of warping factors. The optimal warp factor estimation is given by:

$$\hat{\alpha}_{\mathbf{ML}} = \arg\max_{\alpha} \mathbf{p}\{\mathbf{X}^{\alpha}|\lambda; \mathbf{W}\} \tag{3}$$

where, λ is the HMM model and $\mathbf{W}$ is the true transcription during training and first-pass transcription during recognition. The range of α is usually chosen to be between 0.80 and 1.20 with increments of 0.02. For efficient implementation, VTLN warping is embedded into the the filter-bank and hence the structure needs to be changed for each α before extracting the VTLN warped MFCC features [2].

VTLN can be simplified by deriving a linear transformation on conventional MFCC to obtain VTLN warped MFCC features, i.e.

$$\mathbf{C}^{\alpha} = \mathbf{G}^{\alpha}.\mathbf{C} \tag{4}$$

where, $\mathbf{C}$ and $\mathbf{C}^{\alpha}$ represent the conventional and VTLN warped MFCCs respectively. $\mathbf{G}^{\alpha}$ represents the VTLN matrix for a specific warping factor α. Obtaining such a relation eliminates the need to change the filter-bank structure for each α.

The linear transformation is derived using the idea of band-limited interpolation. The idea is to estimate the VTLN warped mel filter-bank outputs, that are seen as non-uniformly spaced samples given the conventional mel filter-bank outputs, that are seen as uniformly spaced samples. This can be done using band-limited interpolation assuming that the cepstra are band-limited. For more details please refer to [7, 8]. The linear transformation is given by:

$$\mathbf{G}^{\alpha} = [\mathbf{D}.\hat{\mathbf{T}}^{\alpha}.\mathbf{D}^{-1}] \tag{5}$$

where, $\mathbf{D}$ and $\mathbf{D}^{-1}$ are the forward and inverse DCT transformations respectively and $\hat{\mathbf{T}}^{\alpha}$ is the band-limited interpolation matrix for performing spectral scaling. The interpolation matrix is given by:

$$\hat{\mathbf{T}}^{\alpha} = \frac{2}{N-1}\left[\mathbf{U}^{\alpha}_{jk}.\mathbf{V}_{ki}\right] \tag{6}$$

where, N is the number of filters. The matrices $\mathbf{U}^{\alpha}$ and $\mathbf{V}$ are given by:

$$\mathbf{U}^{\alpha}_{jk} = \left[a_k\ cos(2\pi\beta^{\alpha}_j k)\right]_{\substack{0 \le j \le N-1 \\ 0 \le k \le N-1}}, \quad \beta^{\alpha}_j = \frac{\nu^{\alpha}_j}{2\nu_s}$$

$$\mathbf{V}_{ki} = \left[a_i\ cos(2\pi\beta_i k)\right]_{\substack{0 \le k \le N-1 \\ 0 \le i \le N-1}}, \quad \beta_i = \frac{\nu_i}{2\nu_s}$$

and

$$a_i, a_k = \begin{cases} \frac{1}{2}, & i,k = 0, N-1 \\ 1, & i,k = 1,2,\ldots,N-2 \end{cases}$$

$\beta_i, \beta^{\alpha}_j$ are normalized frequencies with the range $0 \le \beta_i, \beta^{\alpha}_j \le 0.5$. ν_i and ν^{α}_j are the mel filter-bank outputs of the conventional and VTLN warped filter-bank respectively and ν_s is the Nyquist frequency in mels. The matrix $\mathbf{G}^{\alpha}$ is derived for static coefficients. It can be easily shown that the same matrix can be used for the delta and acceleration coefficients as well. The matrix $\mathbf{A}^{\alpha}$ is obtained by repeating the matrix $\mathbf{G}^{\alpha}$ in a block diagonal form. The optimal warp factor estimation using the linear transformation is given by:

$$\hat{\alpha}_{\mathbf{ML}} = \arg\max_{\alpha} \mathbf{p}\{\mathbf{A}^{\alpha}\mathbf{X}|\lambda; \mathbf{W}\} \tag{7}$$

Table 1: *Recognition performance (% WER) comparing VTLN and SAT for WSJ0 task.*

	Matched	Mis-matched
Baseline	4.5	31.8
VTLN	3.9	6.5
SAT	3.2	9.8
VTLN + SAT (VS)	2.5	4.2

Here, we store the information of the VTLN matrix that provides the best likelihood rather than storing the warped features.

3. Speaker Adaptive Training

In speaker adaptive training (SAT), the speaker charectaristics are modeled explicitly as linear transformations of the speaker independent (SI) acoustic parameters [9]. CMLLR adaptation matrices are estimated for each speaker by pooling all the data available for a specific speaker both during training and recognition [10]. Since each speaker is transformed separately, speaker variability is accounted by transforming the mean and covariance parameters of the model and are given by:

$$\mu^{(s)} = \mathbf{B}^{(s)}\mu + b \quad \text{and} \quad \Sigma^{(s)} = \mathbf{B}^{(s)}\Sigma\mathbf{B}^{(s)^T} \tag{8}$$

where, $\mathbf{B}^{(s)}$ is the CMLLR transformation matrix for a specific speaker s. Each speaker has its associated transformation matrix that transforms the SI model. More details about CMLLR can be found in [4, 5]. When VTLN is a part of the model building process, the VTLN model will be used as the SI model for estimating the SAT transformations.

4. Recognition Experiments

All the experiments were done using HTK toolkit. We present our analysis on the Wall Street Journal (WSJ0) database, which is wide-band and sampled at 16KHz. The training data consists of 7124 utterances and the test set consists of 330 utterances consisting of both male and female speakers. We perform recognition experiments using both matched and mis-matched train and test speaker conditions. In the matched case, the train and test data include both male and female speakers. Whereas in the mis-matched case, the models are trained using only the male part of the training data (3474 utterances) and tested using only the female part of the test data (123 utterances).

We use triphone HMM models, that consists of 3 emitting states with 16 diagonal covariance mixtures per state. We start with 41 monophones that include silence and short-pause. The triphone models are tied using decision tress and consists of 2259 tied states for the complete data using both male and female speakers and 1383 tied states for the data using only male speakers for training. The features in all the tasks are of 39 dimensions comprising MFCCs appended with delta and acceleration coefficients. In all cases cepstral mean subtraction is applied. We follow a two-pass approach for both VTLN and SAT in recognition.

Table. 1 presents the recognition results comparing Baseline with VTLN, SAT and VTLN+SAT (VS). We observe that SAT performs better than VTLN in the matched case as expected, but VTLN does a good job when compared to SAT in mis-matched case. Since a two-pass approach is followed while estimating the SAT transformations, the transcription errors in baseline might be affecting the performance of SAT. VTLN is considered to be robust to transcription errors and might still

Table 2: *Recognition performance (% WER) using VTLN as SAT transformation matrices in recognition.*

Train	Test	Matched	Mis-matched
No-Adapt	No-Adapt	4.5	31.8
SAT-Train	SAT-Recog	3.2	9.8
SAT-Train	VTLN-Recog	3.6	7.0
SAT-Train	No-Adapt	4.7	37.3

recover. We see that VS gives the best performance in both matched and mis-matched speaker conditions, indicating that including VTLN and SAT in the training process might provide additive improvements and it also indicates that SAT might not completely learn what VTLN can do. In the following sections, we present our ideas on combining VTLN with SAT and how these combinations influence the recognition performance.

4.1. VTLN as SAT Transformation in Recognition

In this section, we present the idea of using VTLN matrices as SAT transformation matrices in recognition. The need for such a combination arises when there is sufficient training data to create a SAT model, but very little adaptation data to estimate a CMLLR transformation matrix in recognition. In such cases, the required adaptation can not be performed because the SAT model expects a speaker transformation for the test data. This forces us to discard the use of SAT in training, even though it might provide a better SI model. It will be advantageous to use the models trained with SAT and estimate a simpler transformation during test with the available adaptation data. In this direction, we present the idea of using VTLN matrices as SAT transformation in recognition because they can be estimated with very little adaptation data.

Table. 2 presents the recognition results for the above discussed combination. The train and test condition have also been specified for better understanding. No Adaptation (No-Adapt) in train and test corresponds to baseline. We observe that, following conventional SAT in training and using the VTLN matrices as SAT transformation in recognition improves the performance when compared with baseline. This is interesting because, VTLN transformed features does not seem to create a mis-match to the model trained using conventional SAT.

In the matched case, we observe that performing VTLN as SAT transformation in recognition does not reach the conventional SAT performance, but has performance better than conventional VTLN (from Table. 1 and Table. 2). In the mis-matched case, we observe that performing SAT in training and VTLN in recognition provides better performance than conventional SAT. The result seems to be interesting, but at this point we are not able explain the reason for this behavior and further investigation is required. The results also include a case where we do not perform any adaptation in recognition but perform SAT in training. From the table it is clear that the performance is inferior compared to baseline, indicating that SAT training is valid only with a transformation on the test data.

4.2. VTLN Before and After SAT

In this section, we present the idea of performing VTLN after SAT. VTLN is always performed before the SAT transformations are estimated or applied. This is because VTLN-scaling is embedded into the mel filter-bank and is a part of the feature extraction. Our motivation to apply VTLN after SAT comes from the fact that VTLN can now be seen as a linear transformation

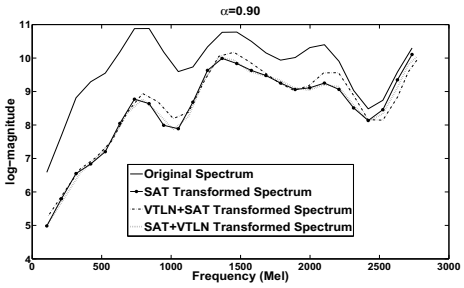

Figure 1: Comparing the effect of various transformations on the log-magnitude spectrum, obtained by performing inverse DCT on the transformed MFCCs.

Table 3: *Recognition performance (% WER) combining VTLN before and after SAT.*

	Matched	Mis-matched
Baseline	4.5	31.8
VTLN	3.9	6.5
SAT	3.2	9.8
VTLN+SAT (VS)	2.5	4.2
SAT+VTLN (SV)	2.6	4.7
VS + VTLN	2.6	3.7

on conventional MFCC. Although this is technically possible, it is difficult to say what comes out when the VTLN matrices are used to transform the features that have been first transformed by SAT.

Figure. 1 shows the effect of SAT, VTLN+SAT (VS) and SAT+VTLN (SV) transformations on the log-magnitude spectrum of a speech signal for a specific speaker. The spectrum is obtained by performing an inverse DCT (IDCT) operation on the MFCC features. In case of SAT, the block matrix estimated from the static coefficients is used to transform the MFCCs and then perform IDCT. Similar operations are performed using VS and SV to obtain the spectrum. It can be seen that, the SAT transformed spectrum tilts when compared with the original spectrum. The spectral scaling by VS is different when compared with SV, with VS showing visible differences when compared with the SAT transformed spectrum. The SV transformed spectrum seems to have not deviated much from the SAT transformed spectrum.

Table. 3 presents the recognition results for different combinations of VTLN and SAT. VS and SV transformations almost perform similarly. For the matched case, it seems like VTLN and SAT operations can be interchanged without affecting the performance. In the mis-matched case, VS performs better than the SV. One of the reasons for this deviation might be the effect of transcription errors in the estimation of SAT matrices. We point out that the VTLN model is better compared to the SAT model in mis-matched case. Since VS provided the best performance in both matched and mis-matched speaker conditions (see Table. 1), we also perform VTLN on top of VS. Performing such a combination helps in the mis-matched case, though it did not affect the performance in matched case.

4.3. VTLN Using Cascaded Transforms

In this section, we present a novel approach to perform VTLN by cascading the VTLN matrices in stages. This is inspired by the idea of using adaptation matrices in cascade as parent and input transforms. The advantage of following such an approach

Table 4: *Recognition performance (% WER) using VTLN matrices in cascade.*

	Matched	Mis-matched
Baseline	4.5	31.8
VTLN	3.9	6.5
VTLN (Iter-1)	4.0	6.6
VTLN-Cascade	3.9	5.7

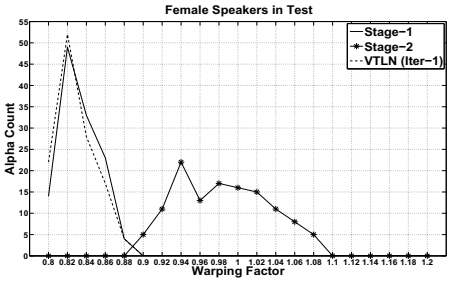

Figure 2: Histogram plots comparing the distribution of warping-factors obtained for the female speaker of the test data at different stages in the cascade. The distribution of warp-factors using the second iteration VTLN model are also presented.

is that, the search range will include warping factors that were not included initially. For example, say that the ML search returned the best warping factor to be 0.90 for a specific speaker. By cascading the VTLN matrices, the warp-factor estimation in the second stage is given by:

$$\hat{\alpha}_{\mathbf{ML}} = \arg\max_{\alpha} \mathbf{p}\{\mathbf{A}_2^{\alpha}\mathbf{A}_1^{0.90}\mathbf{X}|\lambda;\mathbf{W}\} \quad (9)$$

$$= \arg\max_{\alpha} \mathbf{p}\{\mathbf{A}_2^{\alpha}\mathbf{X}^{0.90}|\lambda;\mathbf{W}\} \quad (10)$$

We see that cascading the VTLN matrices will result in scaling of warped features and will result in warp-factors that have not been included in the initial search space.

Table. 4 presents the recognition results using VTLN matrices in cascade. The table also includes the result, where a second iteration of VTLN is performed using the previous iteration VTLN model as the SI model. We observe that, performing iterations over VTLN does not improve the performance in both matched and mis-matched speaker conditions. The proposed approach to cascade the VTLN matrices provides improvement in mis-matched case though there is no effect on the recognition performance in the matched case.

The distribution of warp-factors for the female part of the test data used in mis-matched experiment is presented in Figure. 2. The warp-factor estimates in stage-1 ($\mathbf{A}_1$) will be same as the warp-factors obtained in conventional VTLN. The interesting part is to observe the distribution of warp-factors in stage-2 ($\mathbf{A}_2$). If the warp-factors are not changing from stage-1, then the distribution of warp-factors in stage-2 will be concentrated at 1.00. We see that the distribution of the warp-factors for the female data are spread out from 0.88 to 1.10, indicating that the warp-factors in stage-2 have changed considerably. The figure also includes the distribution of warp-factors using the second iteration VTLN model, which is close to the distribution of stage-1.

5. Conclusion

In this paper, we have presented a variety of experiments to combine VTLN and SAT and showed that performing certain combinations improve the performance of SI-ASR. The main motivation behind these investigations comes from formulating the VTLN-scaling as a linear transformation on conventional MFCC and its interpretation as a pre-computed CMLLR matrix. We showed that VTLN matrices can be used as SAT transformation matrices in recognition when sufficient adaptation data is not available to estimate the CMLLR matrix, though the training still follows conventional SAT. An interesting observation form this experiment is that, VTLN transformed features does not seem to create a mis-match to the models trained using SAT. We showed that the performance in both matched and mismatched speaker conditions improved when compared to the baseline and was close to the conventional VTLN performance. We also presented the idea of performing VTLN after SAT and showed that swapping the order of transformation might have minimal effect on the recognition performance. Finally, we presented a novel approach to perform VTLN by cascading the VTLN matrices. We showed that the search range can include warping factors that have not been part of the initial search. This method proves to be advantageous in mis-matched speaker conditions. Though most of the combinations presented in paper improved the performance of mis-matched speaker conditions, we believe that these combinations might prove to be advantageous in matched speaker conditions as the complexity of the task increases.

Acknowledgement: The research leading to these results was partly funded from the EC in FP7 project EMIME (213845) and from the Academy of Finland in project AIRC.

6. References

[1] A. Andreou, T. Kamm, and J. Cohen Experiments in Vocal Tract Normalization, In *Proc. CAIP Workshop: FSR II*, 1994.

[2] L. Lee and R. Rose. Frequency Warping Approach to Speaker Normalization, *IEEE Trans. SAP*, Vol. 6, pg 49–59, Jan. 1998.

[3] C. J. Leggetter and P.C. Woodland. Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models, *Comp. Speech Lang.*, Vol. 9, pg 171–185, 1995.

[4] V. Digalakis, D. Rtischev, L. Neumeyer, and Edics Sa. Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures, *IEEE Trans. SAP*, Vol. 3, pg 357–366, 1995.

[5] M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition, *Comp. Speech Lang.*, Vol. 12, No. 2, pg 75–98, 1998.

[6] M. J. F. Gales and P. C. Woodland. Mean and Variance Adaptation within the MLLR Framework. *Comp. Speech Lang.*, Vol. 10, pg 249–264, 1996.

[7] D. R. Sanand and S. Umesh. Study of Jacobian Compensation Using Linear Transformation of Conventional MFCC for VTLN, In *Interspeech 2008*, pg 1233–1236, Sept. 2008.

[8] D. R. Sanand, R. Schlüter and H. Ney. Revisiting VTLN Using a Linear Transformation on Conventional MFCC, in *Interspeech 2010*, pg 538–541, Sept. 2010.

[9] T. Anastasakos, J. McDonough, R. Schwartz and J. Makhoul. A Compact Model for Speaker-Adaptive Training, in *ICSLP 96*, pg 1137-1140, Vol. 2, Oct. 1996.

[10] D. Pye and P. C. Woodland, Experiments in Speaker Normalisation and Adaptation for Large Vocabulary Speech Recognition, in *ICASSP 1997*, pg 1047-1050, Vol. 2, Apr. 1997.

[11] C. Breslin, K. K. Chin, M. J. F. Gales, K. Knill and H. Xu, Prior Information for Rapid Speaker Adaptation, In *Interspeech 2010*, pg 1644–1647, Sept. 2010.