IMPROVING THE PERFORMANCE OF VTLN UNDER MISMATCHED SPEAKER CONDITIONS AND MAKING IT APPROACH THAT OF MATCHED SPEAKER CONDITIONS

D. R. Sanand, S. P. Rath and S. Umesh

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

ABSTRACT

The performance of conventional VTLN for mis-matched train and test speaker conditions (e.g. adult-train child-test) does not approach the performance of matched speaker conditions (e.g. child-train child-test). In this paper, we investigate this problem and propose methods to reduce this gap in performance. We use our recently proposed linear transformation approach to VTLN, that also enables us to study the effect of Jacobian unlike conventional VTLN. The main advantage of transform-based VTLN over adaptation based approaches (like CMLLR), is that it does not require any matrix estimation. We argue that the degraded VTLN performance under mismatched speaker conditions is due to the significant frequency warping that is necessary for normalization which leads to a mismatch between the correlation in the feature components of the test data and the covariance structure of the trained/normalized model. We show that the use of a global de-correlating transform (MLLT) leads to improved VTLN performance. We finally show that using both Jacobian and MLLT together improves the VTLN performance for mis-matched cases with the performance approaching that of matched speaker conditions.

Index Terms— Speaker Normalization, VTLN, Linear Transformation, Jacobian, MLLT

1. INTRODUCTION

In this paper, we address the problem where the HMM models are trained on one class of speakers but tested on a significantly different class of speakers. For example, the models may be trained using adult speech data but used for recognizing children speech. This may be necessary since sufficient training data may not be available to build a full fledged model for children. In such cases, there is significant mismatch between train (adult) speakers and test (children) speakers. This leads to significant degradation in recognition performance for children when compared to adults [1, 2]. One of the most commonly used methods to reduce this mismatch is to perform vocal tract length normalization (VTLN) [3]. Although VTLN helps improve the recognition performance for mismatched speakers, there is still a significant gap in performance between the matched and mismatched speaker condition. The goal of this paper is to propose methods to reduce this gap in performance.

Speaker normalization is achieved in VTLN by warping the frequency spectrum of the speech signal, i.e.

$$S_A(f) = S_B(\alpha_{AB} f) \tag{1}$$

A part of this work was supported by SERC project funding SR/S3/EECE/0008/2006 from the Department of Science & Technology, Ministry of Science & Technology, India.

where α_{AB} is the frequency-warp factor used to scale the spectra of speaker B to match the spectra of speaker A. In practice, since there is no reference speaker, a maximum likelihood (ML) based grid search is used to estimate the optimal warp factor α which is given by:

$$\hat{\alpha}_i = \arg\max_{\alpha} \Pr(\boldsymbol{C}_i^{\alpha}|\lambda, W_i)|dC_i^{\alpha}/dC_i| \tag{2}$$

where, C_i^{α} are the feature vectors of the i^{th} utterance frequency-warped by α. λ is the SI or previous iteration VTLN model and W_i is the true transcription in the case of α-estimation during training and the first pass recognition output during testing.The term $|dC_i^{\alpha}/dC_i|$ is the Jacobian of the transformation and it accounts for the mismatch in the likelihood calculation of the warped features C_i^{α} with respect to the SI or previous iteration model (λ). This term is usually ignored in conventional VTLN, since the transformation between warped and unwarped cepstral features cannot be easily determined.

Recently we have shown that, warped features C_i^{α} can be obtained from the conventional MFCC features through a linear transformation, A^{α} [4, 5, 6]. In [4], we discuss the problems of linear transformation approach proposed in [7]. For this paper, we use the method proposed in [6], where the matrix A^{α} is analytically computed for each α and stored. Since, a linear transformation for VTLN can be defined, the Jacobian can be accounted for and is simply the determinant of the transformation. We show that the use of Jacobian helps improve the normalization performance in matched speaker conditions, but degrades the performance in mis-matched speaker conditions. The frequency-warping used in vocal tract length normalization attempts to normalize the spectra of different speakers uttering the same sound. Usually males have longer vocal tracts compared to females and children, resulting in lower frequency formants. In order to match the formant frequencies of a male speaker, the female or child speaker's speech spectra needs to be compressed. We hypothesize in this paper that, significantly warping the frequency spectrum will result in increased correlation in the warped features which may not be modeled well by the covariance structure in the train/normalized model which usually contain a mixture of diagonal covariance Gaussian components. This leads to loss of likelihood when using the warped features of the test data with the trained models, and the use of Jacobian (which assumes proper likelihood calculation) results in over-compensation and hence degradation in performance.

In this paper, we address the issue of correlations in warped features and also study the effect of Jacobian during the transformation in VTLN for the mismatched case. To reduce the correlation, we use a global de-correlating transform like MLLT (maximum likelihood linear transformation) [8, 9]. We show that the use of this decorrelating transform together with the Jacobian will significantly improve the recognition performance in VTLN and also reduce the gap in

performance between matched and mis-matched speaker conditions.

The paper is organized as follows: In Section 2, we review the idea of linear transform for VTLN and discuss how Jacobian is accounted for. We also discuss the effect of Jacobian on the recognition performance. In Section 3, we show the effects of correlations on the recognition performance, especially when there is significant frequency-warping. Later in Section 4, we introduce the idea of a global decorrelating transform that can reduce the correlations in the data. In Section 5, we discuss the experimental setup used in our experiments, followed by results and discussion in Section 6. We finally present our conclusions in Section 7.

2. VTLN USING LINEAR TRANSFORM (LT-VTLN)

VTLN is used in most state of the art speech recognition systems to reduce inter-speaker variability and improve the performance of speaker independent (SI) automatic speech recognition (ASR). VTLN is attractive when compared to transform based adaptation since it requires very little test data to estimate the single frequency warp factor. The main disadvantage with VTLN is that, it requires generation of features for each warp-factor before an optimal estimate of α is found using Eq. (2). We have recently shown that the warped features C^{α} can be generated using a linear transformation (LT) of conventional (un-warped) MFCC features, C [6], i.e.

$$C^{\alpha} = A^{\alpha} C \tag{3}$$

where the warp-matrix A^{α} can be analytically computed given the warping function $g(\alpha, f)$. In this paper, we only consider piecewise linear warping. Apart from generating the warped features using the LT, the Jacobian during α estimation in Eq. (2) can also be accounted. The Jacobian will be simply the determinant of A^{α} in this case.

It is well known that the likelihood of the observed sequence ($\mathbf{C_t}$) with respect to the model parameters ($\mu, \mathbf{\Sigma}$) and the LT matrix ($\mathbf{A}^{\alpha}$) is equivalent to applying the LT on the observation vectors and accounting for the Jacobian of the transformation [10].

$$\mathcal{L}(\mathbf{C_t}; \mu, \mathbf{\Sigma}, (\mathbf{A}^{\alpha})^{-\mathbf{1}}) = \mathcal{N}(\mathbf{A}^{\alpha}\mathbf{C_t}; \mu, \mathbf{\Sigma}) + log(|\mathbf{A}^{\alpha}|) \tag{4}$$

Note that this is similar to the CMLLR approach to adaptation, except that, in this case *the VTLN matrix, A^{α}, is already precomputed* for each α or equivalently we can transform the features and account for the Jacobian. Similar to conventional VTLN, we estimate the warp-factor for each training utterance, normalize the features and estimate a normalized model. This is repeated a few iterations. During testing , we choose the appropriate pre-computed matrix A^{α} to obtain normalized features of the test data and do the final recognition.

Table 1 and Table 2 show the recognition performance for matched and mis-matched cases respectively using our recently proposed linear-transformation (LT) approach [6]. We have shown that the performance of conventional and the proposed LT-VTLN (with no Jacobian) give comparable performance in [6]. The details of the experimental set up and databases will be discussed in detail in Section 5. These tables show the normalization performance with and without the use of Jacobian. From the tables, we make the following observations:

- Accounting for Jacobian has provided improved recognition performance on all tasks in matched cases.
- Accounting for Jacobian has degraded the recognition performance on all tasks in mis-matched cases.

VTLN performs speaker normalization by scaling the frequency spectrum of the speech signal. The scaling/warping factor is estimated with respect to the SI or the previous iteration VTLN model using Eq. (2). The warp factor is restricted to be in the search range from 0.80 to 1.20 based on physiological arguments and usually increments of 0.02 are followed. When there is significant mis-match between the train and test sets the warp factors are usually far from unity, i.e. for example, if we use male data to train the model and are using it to recognize children speech, all the warp factors for child speakers will be close to 0.80.

Table 1. Recognition Performance of LT-VTLN With and Without Jacobian Compensation for Matched Cases

Method	TIDIGITS		RM-Task	OGI
	F-F	C-C	A-A	A-A
Baseline (No-VTLN)	99.79	99.71	96.49	96.95
LT-VTLN	99.79	99.62	96.92	97.40
LT-VTLN + JACOB	99.81	99.75	97.07	97.64

- A-A – Adult train - Adult test
- F-F – Female train - Female test
- C-C – Child train - Child test

Table 2. Recognition Performance of LT-VTLN With and Without Jacobian Compensation for Mis-matched Cases

Method	TIDIGITS		OGI
	M-F	M-C	A-C
Baseline (No-VTLN)	94.52	69.08	86.20
LT-VTLN	99.49	96.20	92.91
LT-VTLN + JACOB	99.08	87.50	90.93

- A-C – Adult train - Child test
- M-F – Male train - Female test
- M-C – Male train - Child test

We hypothesize that, significant frequency warping for performing speaker normalization will introduce correlations in the components of the warped features. These correlations are not modeled well by the covariance structure in the trained/normalized model. This mismatch results in the loss of likelihood when mis-matched speakers are tested against this model. Therefore, in Table 2, the use of Jacobian (which assumes proper likelihood calculation) results in *over-compensation* leading to degradation in performance.

In order to understand the reason for this degradation in mismatched cases after accounting for Jacobian, we perform a set of experiments to understand whether correlations introduced due to warping have any role. These experiments are done in an adaptation frame-work and are discussed in more detail in the next section.

3. ADAPTATION EXPERIMENTS TO UNDERSTAND EFFECT OF CORRELATIONS IN WARPED FEATURES

In CMLLR adaptation [10], both the means and covariances are transformed by the *same* matrix, and is equivalent to a feature transformation as shown in Eq. 4. We argue that while the transformation of the mean helps bring the model closer to the mismatched data, the use of the same transformation for transforming the covariance of the model results in a mismatch between the covariance of the feature and model components. Note that there is an implicit Jacobian compensation in CMLLR as seen in Eq. 4. Unlike VTLN-matrix, in this case the CMLLR matrix is estimated in a ML frame-work. We also show the adaptation results for MLLR - MEAN [11] (the estimated adaptation matrix is used to transform only the means, the variances are un-affected) and MLLR - MEAN+VAR [12] (estimates a separate adaptation matrix for both means and variances). The recognition results for these various transformation approaches are presented in Table 3. We use two different set of mis-matched cases from TIDIGITS, namely male train - female test (M-F) and

Table 3. Recognition Performance of Various Adaptation Based Approaches for Mis-matched Cases on TIDIGITS

Method	TIDIGITS	
	M-F	M-C
Baseline (No-VTLN)	94.52	69.08
CMLLR	99.42	92.85
MLLR - MEAN	99.49	95.18
MLLR - MEAN+VAR	99.51	95.58

male train - child test (M-C). We observe that:

• In both the mismatched cases, CMLLR has inferior recognition performance compared to other adaptation based approaches. Applying a transformation on the test data features might result in components being correlated. These correlated features may not be properly modeled by the covariance matrices of the train model (and Jacobian over-compensation may occur). This mismatch results in performance degradation.

• In the case of MLLR - MEAN the variances are the same as in the original train model and only the means of the adapted model are transformed. We argue that there is less mismatch in the correlation structure with respect to the train model and hence the recognition performance is better than CMLLR.

• In case of MLLR - MEAN+VAR, there is a different transformation for both means and variances. We observe that providing this freedom has provided gain over both the other adaptation based approaches. This is because, the covariance structure of the adaptation data are also now properly modeled using a separate matrix.

These results show that using a single transformation matrix for adapting both means and variances might result in degraded performance in the mis-matched cases. This is because the Jacobian over-compensates the likelihood. This also explains the reason for degraded performance of VTLN in the LT framework when Jacobian is accounted in mis-matched cases in Table 2.

In order to address the problem of differences in correlation structure in the mismatched case, we try to estimate a global de-correlating transform like MLLT using the trained model and small amount of train data (different from test set) from mismatched speaker set. This will help reduce the mismatch in the covariance structure and our experiments, shown later in the paper, indicate improved performance. In the next section, we briefly review MLLT.

4. BRIEF REVIEW OF MLLT

Maximum Likelihood Linear Transformation (MLLT) [8] is a special case of Heteroscedastic Linear Discriminant Analysis (HLDA) [9]. HLDA is a class discriminant analysis method that does not assume equal covariance between the classes, as opposed to Linear Discriminant Analysis (LDA). It finds a projection of the n dimensional feature vectors to a new space and constrains all the classes in the projected space to have the same mean and covariance in the last $n-p$ ($p < n$) components. The first p components in the new space have different mean and covariance among the classes. Hence the class discrimination is contained only in the first p components. The projecting transform is obtained using maximum likelihood criterion. Effectively HLDA transforms the n dimensional features space to a p dimensional space and thus reduces the dimensionality from n to p.

When HLDA is configured with $p = n$, i.e., when there is no dimensionality reduction it is known as MLLT. In this paper we have used the case of MLLT where diagonality constraint is forced on the covariance matrices in the projected feature space so that MLLT works like a decorrelating transform for the feature vectors. MLLT aims at minimizing the loss in likelihood between full and diagonal covariance Gaussian models.

5. RECOGNITION EXPERIMENTS

The recognition experiments are performed on three different databases, which include Resource Management (RM) task, TIDIGITS and Number corpus of OGI. RM and TIDIGITS are wide-band speech having a sampling frequency of 16KHz and 20KHz respectively. OGI is a narrow-band telephone based continuous digit corpus with a sampling frequency of 8 KHz. RM is an adult-speaker database consisting of 3990 utterances for train and 300 for test. TIDIGITS consists of males (4235 for train, 4311 for test), females (4388 for train, 4388 for test) and children(3925 for train and 3847 for test), whereas OGI also consists of adults (6078 for train, 2169 for test) and children (2798 for test) data. Based on this we formulate different combination of experiments for matched and mis-matched conditions.

In TIDIGITS and OGI, the digits are modeled as whole word simple left-to-right HMMs without skips and have 16 states per word with 5 diagonal covariance Gaussian mixtures per state. On the RM database we perform the recognition task using state-tied cross-word triphones. We use phonetic decision tree based clustering for tying the states. The phone HMM models consist of 3 states with 6 diagonal covariance Gaussian mixtures per state. In both the tasks, we used a silence model having 3 states and a single state short pause model tied to the middle state of the silence model. The features in all tasks are of 39 dimensions comprising normalized log-energy, $c_1, \ldots, c_{12}$(excluding c_0) and their first and second order derivatives. In all cases, cepstral mean subtraction was applied. For performing VTLN, only un-warped features are generated and all the warped features are generated on the fly using the linear transform matrices as shown in Eq. 3. Note that the matrices for different warp-factors, α, are pre-computed for the specific warping function.

6. RESULTS AND DISCUSSION

In Table 4, we present results on two combination of experiments namely matched and mis-matched speaker conditions. For the matched case, the experiments include female train - female test (F-F) and child train - child test (C-C) from TIDIGITS, adult train - adult test (A-A) from RM task and adult train - adult test (A-A) from OGI. For mis-matched conditions, the experiments include male train - female test (M-F) and male train - child test (M-C) from TIDIGITS and adult train - child test (A-C) from OGI.

We have already discussed the results about Jacobian compensation in Section 2, where we have shown that Jacobian compensation provides improved recognition performance for matched speaker conditions but provides degradation in performance for mis-matched speaker conditions. We have argued in Section 3 that the degradation in performance is due to differences in correlation structure between matched and mismatched data which results in Jacobian providing over-compensation during likelihood calculation. The use of a global de-correlating transform reduces this mismatch in correlation structure and hence the use of subsequent Jacobian compensation should help improve the recognition performance. We present the results for both matched and mis-matched cases after applying MLLT followed by Jacobian compensation.

We make the following observations in matched cases:

• Applying MLLT without Jacobian (LT-VTLN+MLLT) has provided improvement over LT-VTLN.

• Accounting for Jacobian after MLLT (LT-VTLN + MLLT + JACOB) has provided additional gain in the performance.

Table 4. Recognition performance of the proposed LT with and without Jacobian Compensation as well as with and without MLLT for both matched and mis-matched speaker conditions.

Method	Matched				Mis-matched		
	TIDIGITS		RM-Task	OGI	TIDIGITS		OGI
	F-F	C-C	A-A	A-A	M-F	M-C	A-C
Baseline (No-VTLN)	99.79	99.71	96.49	96.95	94.52	69.08	86.20
LT-VTLN	99.79	99.62	96.92	97.40	99.49	96.20	92.91
LT-VTLN + JACOB	99.81	99.75	97.07	97.64	99.08	87.50	90.93
LT-VTLN + MLLT	99.79	99.68	97.54	97.69	99.74	98.98	93.73
LT-VTLN + MLLT + JACOB	99.82	99.79	97.62	97.99	99.79	99.31	93.73

We make the following observations in mis-matched cases:
• Applying MLLT without Jacobian has provided improvement over LT-VTLN.
• Accounting for Jacobian after MLLT has provided additional gain in the recognition performance except for the case of OGI where it remains same. This is in contrast to the performance when Jacobian compensation is used without MLLT.
• The recognition results of matched TIDIGITS F-F and C-C are almost similar or close to the mis-matched cases of TIDIGITS M-F and M-C in the case of LT-VTLN + MLLT + JACOB.

We could not perform the matched case experiment in OGI as there was no training data available for children. Note that since OGI is telephone-speech data (and bandlimited to less than 3400 Hz), children will have poorer performance than adults.

The results indicate that, if we can reduce the differences in the correlation structure in warped features of test data and train model (using MLLT) and also account for the Jacobian during α estimation in VTLN, the recognition performance can be significantly improved. In this case, the performance of mis-matched cases in TIDIGITS is close to the matched cases. In all the cases, we observe that the recognition performance has improved significantly over VTLN after applying a global decorrelating transform and accounting for Jacobian, indicating that both play a major role. We point out the case of M-C in TIDIGITS, that has shown significant improvements at every stage and is also the extreme case of mis-match among all the combinations considered in this paper. We observe that, after the application of MLLT and Jacobian, the performance of M-C approaches the case of C-C.

7. CONCLUSION

In this paper, we have proposed methods to improve the performance of VTLN in mis-matched cases. We have used our recently proposed linear transform approach to VTLN which allows us to study the effect of Jacobian on the warp factor estimation. We showed through experiments that, although matched cases show improvements with Jacobian compensation, the mis-matched cases have degraded recognition performance after Jacobian compensation. We hypothesized and corroborated with the experiments that the significant frequency warping in mis-matched cases increased the correlations in the feature components of the warped features which lead to a mismatch with the covariance structure of the trained model. This leads to an underestimation in the likelihood calculation and hence the use of Jacobian (which assumes proper likelihood calculation) results in over-compensation of the likelihood resulting in performance degradation in mis-matched cases. We have proposed the use of a global de-correlating transform, MLLT, that would reduce the mis-match in the covariance structure between test data and train model. We show that the use of MLLT indeed improved the recognition performance both in matched and mis-matched speaker conditions. But more importantly the use of Jacobian in addition to MLLT gives rise to further improvement in the recognition performance in the mis-matched cases also. Further, the performance of mis-matched cases are now close to matched speaker case. Therefore, we conclude that taking care of the correlations introduced by warping the data and accounting for the Jacobian during warp factor estimation will provide significant gains in the recognition performance over VTLN.

8. REFERENCES

[1] D. Giuliani and M. Gerosa, "Investigating Recognition of Children's Speech," in *Proc. IEEE ICASSP 2003*, vol. 2, Trento, Italy, April 2003, pp. 137–140.

[2] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic Variability and Automatic Recognition of Childrens Speech," *Speech Communication*, vol. 49, no. 10-11, pp. 847–860, Nov 2007.

[3] L. Lee and R. Rose, "Frequency Warping Approach to Speaker Normalization," *IEEE Trans. SAP*, vol. 6, pp. 49–59, Jan 1998.

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

[5] D. R. Sanand, D. D. Kumar, and S. Umesh, "Linear Transformation Approach to VTLN Using Dynamic Frequency Warping," in *Interspeech2007*, Belgium, Sep 2007.

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

[7] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Trans. ASLP*, vol. 13, no. 5, pp. 930–944, 2005.

[8] R. A. Gopinath, "Maximum Likelihood Modeling with Gaussian Distribution for Classification," in *Proc. IEEE ICASSP 1998*, vol. 2, May 1998, pp. 661–664.

[9] N. Kumar, "Investigation of Silicon-auditory Models and Generalization of lda for Improved Speech Recognition," Ph.D. dissertation, John Hopkins Univ., 1997.

[10] M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[11] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden markov Models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[12] M. J. F. Gales and P. C. Woodland, "Mean and Variance Adaptation within the MLLR Framework," *Computer Speech and Language*, vol. 10, no. 4, pp. 249–264, 1996.