INTERSPEECH 2009 BRIGHTON

A Study on the Influence of Covariance Adaptation on Jacobian Compensation in Vocal Tract Length Normalization

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

Department of Electrical Engineering,
Indian Institute of Technology Kanpur, Kanpur, India - 208 016
[drsanand,srath,sumesh]@iitk.ac.in

10.21437/Interspeech.2009-206

Abstract

In this paper, we first show that accounting for Jacobian in Vocal-Tract Length Normalization (VTLN) will degrade the performance when there is a mismatch between the train and test speaker conditions. VTLN is implemented using our recently proposed approach of linear transformation of conventional MFCC, i.e. a feature transformation. In this case, Jacobian is simply the determinant of the linear transformation. Feature transformation is equivalent to the means and covariances of the model being transformed by the inverse transformation while leaving the data unchanged. Using a set of adaptation experiments, we analyze the reasons for the degradation during Jacobian compensation and conclude that applying the same VTLN transformation on both means and variances does not fully match the data when there is a mismatch in the speaker conditions. This may have similar implications for constrained-MLLR in mismatched speaker conditions. We then propose to use covariance adaptation on top of VTLN to account for the covariance mismatch between the train and the test speakers and show that accounting for Jacobian after covariance adaptation improves the performance.

Index Terms: Jacobian Compensation, Linear Transformation, VTLN, Covariance Adaptation, Speaker Normalization, Automatic Speech Recognition.

1. Introduction

Inter speaker variability is a major source of performance degradation in speaker independent (SI) automatic speech recognition (ASR) systems. Inter speaker variability arises due to a variety of speaker specific characteristics, but is mainly attributed to the variations in the vocal tract length (VTL). There is about 25% variability in the VTL across adult male and adult female speakers, with males having the largest and children having the smallest VTL's. Vocal tract length normalization (VTLN) is a standard procedure for performing speaker normalization is SI-ASR systems [1].

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

$$S_R(f) = S_T(\alpha_{RT} f) \tag{1}$$

where α_{RT} is the frequency-warp factor used to scale the spectra of speaker T to match the spectra of speaker R. 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 [2]

$$\Pr(\mathbf{X}_i|\mathbf{W}_i;\lambda^\alpha) = \Pr(\mathbf{X}_i^\alpha|\mathbf{W}_i;\lambda).\,|d\mathbf{X}_i^\alpha/d\mathbf{X}_i| \tag{2}$$

where λ is the SI model, $\mathbf{W}_i$ is the known transcription and $\mathbf{X}_i^\alpha$ is the feature vector of i^{th} utterance whose spectra is warped with the scale factor α. During recognition, $\mathbf{W}_i$ is obtained from the first recognition pass. Since the likelihood of the warped utterance is evaluated with respect to the SI (or previous iteration VTLN) model, the Jacobian of the transformation needs to be accounted for proper likelihood calculation and is given by $|d\mathbf{X}_i^\alpha/d\mathbf{X}_i|$. However, the transformation between cepstral coefficients is difficult to determine, and therefore, the Jacobian of the transformation is usually ignored in practice.

In the recent past there has been lot of interest in finding a linear transformation to obtain VTLN warped cepstral features given the un-warped cepstral features. The use of such a linear transformation makes VTLN computationally efficient since it does not require the computation of warped features for all warp factors before estimating the optimal warp factor in Eq. 2. Another advantage of obtaining a linear transform relation is that we can also study the effect of Jacobian on VTLN performance, which theoretically needs to be taken into account for proper likelihood calculation. The Jacobian in this case will be simply the determinant of the linear transformation.

There have been very few results reported on the study of Jacobian in VTLN. Pitz [3] reported in his thesis that Jacobian provided marginal improvements. Panchapagesan [4] has reported that Jacobian degraded the performance and hence was ignored during recognition. There have been other linear transformation approaches like McDonough [5], Claes [6] and Ciu-Alwan [7] but none have reported any study using the Jacobian.

In this paper, we address the problem of accounting for Jacobian in VTLN using our recently proposed linear transformation on conventional MFCC features [8]. We observe that Jacobian behaves very differently based on the train and test speaker conditions. Jacobian compensation provides improvement in performance for matched speaker conditions and degradation in performance for mismatched speaker conditions. We conjecture that these variations may be due to mismatch in the covariance structure between the trained model and the test speakers. We propose to use covariance adaptation on top of VTLN and show that it improves recognition performance irrespective of the variations in train and test conditions.

The paper is organized as follows. In Section 2, we study the effect of Jacobian and show the performance on matched and mismatched speaker conditions. In Section 3, using a set of adaptation experiments we show that mismatch in the covariance structure might be a possible source of degradation when Jacobian is accounted for. In Section 4, we propose a method of covariance adaptation in the VTLN frame work. In Section 5, we explain the details of our experiments. In Section 6, we present the results obtained using the proposed ap-

Table 1: *Recognition Performance of LT-VTLN With and Without Jacobian Compensation*

Method	TIDIGITS M-C	RM-Task A-A
Baseline (No-VTLN)	69.08	96.60
LT-VTLN (No-Jacob)	96.45	97.07
LT-VTLN (Jacob)	87.37	97.07

• A-A - Adult train - Adult test • M-C - Male train - Child test

proach to show that Jacobian improves the performance when properly accounted for in VTLN. In Section. 7 we present our conclusions followed by references.

2. Study of Jacobian in VTLN

We have recently shown that warped cepstral features X^α can be generated using a linear transformation (LT) of conventional (un-warped) MFCC features, X [8], i.e.

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

where the warp-matrix A^α can be analytically computed given the warping function $g(\alpha, f)$. Obtaining such a relation will enable us to study the effect of Jacobian, as it will be simply the determinant of the transformation matrix A^α.

Before proceeding further, we present the results showing the effect of Jacobian on VTLN performance. The experiments are performed in a manner similar to the conventional filter bank based VTLN, with the only difference being in choosing the appropriate warping matrix. The details of the the experimental setup are described later in Section 5.

Table. 1 shows recognition results with and without Jacobian compensation using the proposed linear transformation. The experiments are performed on TIDIGITS (male train – children test) and RM1 (adult train – adult test) databases. It can be observed from the table that in case of M-C, performance degrades after taking Jacobian into account in the warp factor estimation. In case of A-A, there is no improvement in recognition performance after Jacobian compensation. We believe that although Jacobian compensation might not provide huge improvements, it should improve recognition performance when properly accounted for. It is therefore surprising to note that M-C performance drops drastically, indicating that Jacobian is not taken into account correctly.

We believe that the reason for this degradation is the improper calculation of likelihood while estimating the optimal warp factor using Eq. 2. This can happen when there is a mismatch between the model and the warped features. If the likelihood calculation is not proper, Jacobian can over compensate the likelihood and hence result in the degradation of recognition performance. In Eq. 3, the mismatched data X may be assumed to come from a distribution with model parameters μ_T and Σ_T. Applying VTLN linear transform A^α to get warped features X^α would imply that X^α has model parameters $A^\alpha\mu_T$ and $A^\alpha\Sigma_T(A^\alpha)^T$. These model parameters should match the parameters of the SI model (or previous iteration VTLN model) since the data is warped to match this model. If mismatched data X is indeed *exactly* related to the model observations through a linear transformation, then the likelihood calculation will be perfect and Jacobian compensation is mathematically correct. However, we conjecture that warping the data might match the means properly, but there might be differences in the covariance structure between the model and the data. This is because VTLN-warping is done to match the spectra and therefore the means should match, since the features are derived from the spectra. However, the covariance structure need not match. From the results in Table. 1, we also observe that without Jacobian there is a huge gain in VTLN performance indicating a reasonable but not perfect match to the model. In the next section, we perform a set of adaptation experiments to understand whether the mismatch in the covariance structure might be responsible for performance degradation when Jacobian is accounted.

Table 2: *Recognition Performance of Various Adaptation Experiments*

Method	TIDIGITS M-C	RM-Task A-A
Baseline (No-VTLN)	69.08	96.60
CMLLR	91.56	96.92
MLLR (mean-only)	94.95	96.88
MLLR (mean + cov)	95.01	97.27

• A-A - Adult train - Adult test • M-C - Male train - Child test

3. Mismatch in Covariance Structure

We perform a set of adaptation experiments, like CMLLR, mean-only MLLR and separate MLLR mean and covariance adaptation in order to understand the effect of mismatch in the covariance structure. The reason for performing these experiments is to understand how each of these transformations affect the performance for matched and mismatched train and test speaker conditions. The results are shown in Table. 2

In all the above experiments, we derive a separate adaptation matrix for each speaker on the test data using the baseline transcription. In the case of MLLR (mean + cov) for TIDIGITS, we performed diagonal covariance adaptation (MLLRVAR in HTK[9]) as the data was not sufficient for deriving full covariance adaptation (MLLRCOV in HTK [9]).

For the mismatched train and test speaker conditions (M-C), we observe that CMLLR has a performance much inferior to Mean-only adaptation. This is not the case in matched speaker conditions (A-A), where it is slightly better than Mean-only adaptation. We observe that in both the tasks, using a separate transformation for means and covariances provides the best performance.

CMLLR can be seen as an equivalent to VTLN with Jacobian compensation, as it can also be viewed as a feature domain transformation. This is to say that the likelihood of the observed sequence ($\mathbf{X_t}$) with respect to the model parameters (μ, $\mathbf{\Sigma}$) and the transformation matrix ($\mathbf{B}$) is equivalent to applying the inverse transformation matrix on the observation vectors and accounting for the Jacobian of the transformation [10].

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

In our approach to linear transformation for VTLN, we have a separate transformation matrix $\mathbf{A}^\alpha(= \mathbf{B^{-1}})$ for each warp factor which is used in generating warped features. The variation of *logdet* values for different warp factors is shown in Fig. 1.

From the above experiment, we conclude that using a single transformation matrix to transform both means and variances degrades the performance in mismatched speaker conditions. In other words, applying a single transformation matrix on both the means and the covariances will not provide a complete match between the data and the model.

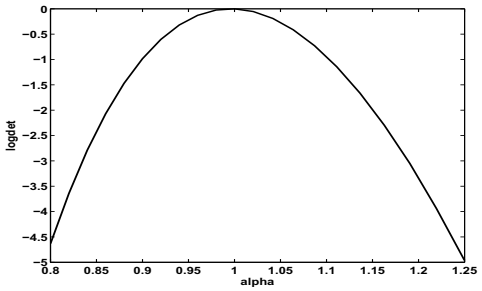

Figure 1: *Plot showing the variation of* $\log(|\mathbf{A}^{\alpha}|)$ *for RM task.*

In the case of VTLN, presuming that frequency-warping to normalize the spectra is able to match the means properly, performing some sort of covariance adaptation might improve the performance when Jacobian is accounted in VTLN. In the next section, we propose a method of covariance adaptation in the VTLN frame work to reduce the mismatch between the trained model and the test speakers.

4. Covariance Adaptation

As discussed in the previous sections, mismatch in the covariance structure might be responsible for degradation in performance when Jacobian is accounted. We propose to perform covariance adaptation on top of VTLN to overcome this effect. This implies that we keep the means unchanged and adapt the covariances of the trained model to better represent the covariance of the test data.

For example, let us consider the case of male train and children testing (case of M-C). Let λ_M represent the male trained model, with μ_M and Σ_M representing the means and covariances respectively. Now we estimate a new model using child (adaptation) data to adjust the covariance structure of the male model to better represent the children data. Mathematically,

$$\mu_M = \mu_M \quad \text{and} \quad \hat{\Sigma_M} = H\Sigma_M H^T \tag{5}$$

where $\hat{\Sigma_M}$ is the modified covariance and H represents the covariance transformation. The steps in creating the covariance adapted model can be summarized as:

- Create the male train model using the male training data (λ_M).
- Now using children data (adaptation data), perform alpha estimation *without* Jacobian. We follow this because, initially accounting for Jacobian is not proper.
- Now using the alpha estimation from the previous step to obtain warped features and λ_M, we create a variance adapted model, say λ_{MVNJ}, (here M- Male, V- Variance Adapted NJ - No Jacobian).
- Using λ_{MVNJ}, we now perform alpha estimation on the adaptation data *with* Jacobian.
- Now create a new variance adapted model (λ_{MVJ}) using λ_{MVNJ} and the warped data based on alpha estimation from the previous step.

The procedure is illustrated in Fig. 2.

5. Recognition Experiments

The recognition experiments are performed on two different databases, which include Resource Management (RM1) task

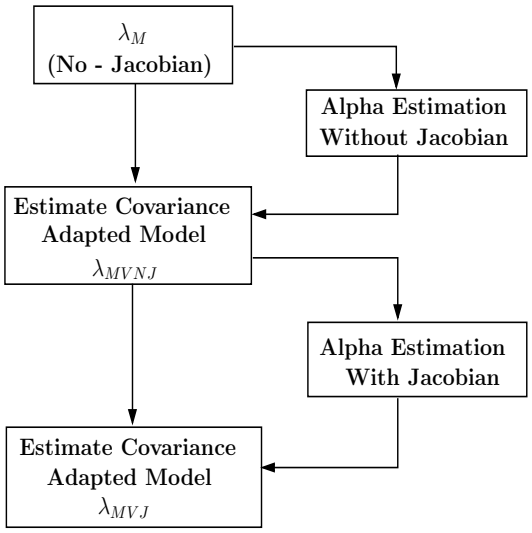

Figure 2: *Illustrating the training procedure for creating the covariance normalized model*

and TIDIGITS. Both the databases are wide-band speech having a sampling frequency of 16KHz for RM1 and 20KHz for TIDIGITS respectively. RM1 is an adult-speaker database consisting of 3990 utterances for train and 300 for test. TIDIGITS consists of 4235 utterances for male train and 3847 utterances for children testing.

In TIDIGITS, 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 RM1 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.

VTLN is performed both during training and testing and we perform warp factor estimation at the utterance level. While performing recognition in VTLN, we follow a two pass approach. Jacobian is compensated both during training and testing. While performing the covariance adaptation experiments, we use the train part of the children data (3925 utterances) available for TIDIGITS to find the covariance adaptation matrix. In case of RM task, since the train and test conditions are matched we use the train data itself to estimate the covariance adaptation matrix and do not use any information from the test data.

6. Results and Discussion

Table. 3 presents the results using the proposed method for speaker normalization for VTLN using Jacobian compensation. We observe that after performing covariance adaptation and then accounting for Jacobian, there is no degradation in performance. It has improved the performance over the no-Jacobian case as well. We also observe that the performance for RM task has improved. We believe that Jacobian will not provide big improvements but when accounted properly should not degrade the performance.

Table 3: *Recognition Performance Comparing VTLN and Covariance Adapted model with Jacobian Compensation*

Method	TIDIGITS M-C	RM-Task A-A
Baseline (No-VTLN)	69.08	96.60
LT-VTLN (No-Jacob)	96.45	97.07
LT-VTLN (Jacob)	87.37	97.07
Cov Adaption (Jacob)	97.12	97.27

• A-A - Adult train - Adult test • M-C - Male train - Child test

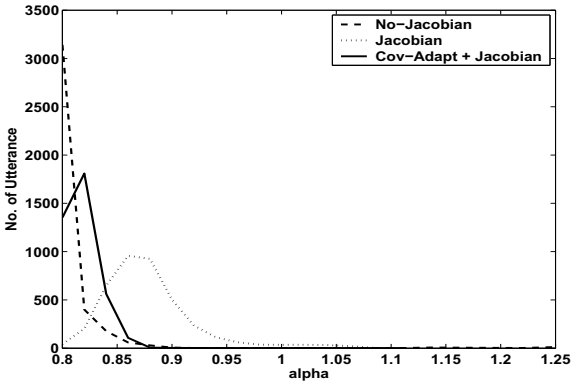

Figure 3: *Histogram plot of the alpha estimates on the TIDIGITS test data before and after Jacobian compensation as well as after covariance adaptation*

Fig. 3 and Fig. 4 show the histogram plots for TIDIGITS and RM test data respectively before and after Jacobian compensation as well as after performing the covariance adaptation. We observe that the alpha estimates for the TIDIGITS task with Jacobian compensation are concentrated around 0.86 and 0.88, where as in the no Jacobian case around 0.80. The alpha estimates after covariance adaptation and taking into account the Jacobian are concentrated around 0.82. For the case of RM we observe that the alpha estimates for the covariance adapted model lie in between the no-Jacobian and Jacobian cases.

7. Conclusion

In this paper, we have proposed a method to properly account for the Jacobian in VTLN. We showed that the performance of VTLN degrades after accounting for Jacobian when there is a mismatch between the train and test speaker conditions. We performed a set of adaptation experiments that suggest that the mismatch in the covariance structure might be responsible for this degradation. We then proposed a covariance adaptation approach in the VTLN framework to better match the covariance of the trained model with the test data. We showed that using the proposed scheme along with Jacobian compensation provides the best performance both in matched and mismatched speaker conditions.

8. Acknowledgment

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.

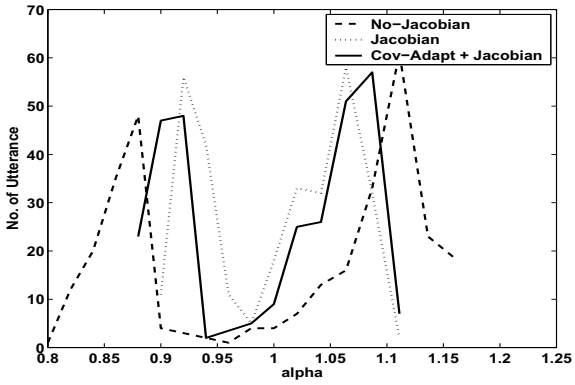

Figure 4: *Histogram plot of the alpha estimates on the RM test data before and after Jacobian compensation as well as after covariance adaptation*

9. References

[1] Li Lee and R. C. Rose, " A Frequency Warping Approach to Speaker Normalization", *IEEE Trans. on Speech and Audio Process.*, Vol. 6, No. 1, pp.49-60, Jan. 1998.

[2] A. Sankar and C.H. Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition", *IEEE Trans. on Speech and Audio Process.*, Vol. 4, No. 3, pp. 190202, May 1996.

[3] Michael Pitz, "Investigations on Linear Transformations for Speaker Adaptation and Normalization", *PhD thesis*, RWTH Aachen University. 2005.

[4] S. Panchapagesan and A. Alwan, "Frequency Warping for VTLN and Speaker Adaptation by Linear Transformation of Standard MFCC", *Computer Speech and Language*, Vol. 23, No. 1, pp. 4264, Jan 2009.

[5] J. McDonough, T. Schaaf, and A. Waibel, "Speaker Adaptation with All-Pass Transforms", *Speech Communication*, Vol. 42, No. 1, pp. 7591, Jan. 2004.

[6] T.Claes, I.Dologlou, L.ten Bosch, D. van Compernolle, "A novel feature transformation for vocal tract length normalisation in automatic speech recognition", *IEEE Trans. on Speech and Audio Proc.*, Vol.6, pp. 549-557, Nov. 1998.

[7] X. Cui and A. Alwan., "Adaptation of childrens speech with limited data based on formant-like peak alignment", *Comp. Speech & Lang.*, Vol.20, Oct. 2006, pp. 400-419.

[8] D. R. Sanand and S. Umesh, "Study of Jacobian Compensation Using Linear Transformation of Conventional MFCC for VTLN", *Interspeech2008*, pp: 1233-1236, Brisbane, Australia, Sept. 2008.

[9] S. Young, G. Evermanm, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book (for HTK Version 3.4). Cambridge, UK, Eng. Dept., Cambridge Univ., 2009.

[10] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition", *Computer Speech and Language*, Vol. 12, No. 2, pp. 7598, April. 1998.