Effect of Jacobian Compensation in Linear Transformation based VTLN under Matched and Mis-matched Speaker Conditions

S. P. Rath, A. K. Sarkar and S. Umesh
Indian Institute of Technology Madras
Email: shaktirath@gmail.com, sarkar.achintya@gmail.com, umeshs@iitm.ac.in

***Abstract*—In this paper we study the effect of use of jacobian in different linear transformation (LT) based methods of VTLN. In conventional VTLN, the jacobian is highly non-linear and can not be computed and hence is ignored. In the LT based VTLN, since VTLN scaling is expressed as a matrix multiplication of un-warped MFCC features, jacobian is simply turns out as the determinant of the VTLN warp matrices. Hence in this framework of VTLN it is possible to account for jacobian. Two different methods, namely, L-VTLN and T-VTLN, are used for implementing LT based VTLN. By conducting experiments on the RM task and the TIDIGITs databases in matched and mis-matched speaker conditions, the performance of using jacobian in warp-factor estimation have been evaluated. It is observed that in almost every matched and mis-matched speaker conditions jacobian improves performance in L-VTLN framework. In T-VTLN, however, jacobian does not improve the performance in any mis-matched speaker conditions. The cases in which jacobian degrades performance in L-VTLN and T-VTLN have been studied in detail.**

I. Introduction

One of the major reasons for the inferior performance of Speaker Independent (SI) Automatic Speech Recognition (ASR) as compared to the Speaker Dependent (SD) ASR is the differences in the Vocal Tract Length (VTL) among speakers. Due to this variations in the VTL, the spectra of the same sound when spoken by different speakers vary considerably and hence the MFCC acoustic features that are extracted from the speech spectrum also vary considerably between speakers. Since, the SI model is trained using the speech utterances collected over many speakers, it captures the spectral variations of the speakers in the training set. As a result, the SI acoustic model do not exactly match the speech characteristics of the test speakers. This causes the performance of SI systems to fall below that of SD systems.

To improve the performance of SI systems and to make it approach to that of the SD models, methods of speaker normalization are used. One of the methods for speaker normalization is called Vocal Tract Length Normalization (VTLN). VTLN tries to reduce the variation in the spectra due to differences in the VTL between speaker by scaling the frequency axis of the spectrum (section II for details). In MFCC processing [1], VTLN is usually implemented by scaling the frequency axis (or, equivalently, scaling the filter-bank in MFCC processing) of the speech spectrum to produce the VTLN warped features. For every utterance the frequency warp-factor is estimated such that the variations in spectrum in the speech utterance is reduced. Since there is no reference speaker to which the spectra of the speech utterance can be matched, usually Maximum Likelihood (ML) is used to obtain the warp-factor, i.e.,

$$\alpha^* = \arg\max_{\alpha} \left\{ log\, p(X^{\alpha}|\lambda, U) \;+\; log \left| \frac{dX^{\alpha}}{dX} \right| \right\} \qquad (1)$$

where, λ, U and $\left|\frac{dX^{\alpha}}{dX}\right|$ are the acoustic model parameter, the transcription and the *Jacobian*, respectively. Also, in Eq.1, X represents the un-warped (No-VTLN MFCC features) and the corresponding MFCC features obtained after warping the spectrum by α is denoted by X^{α}.

Due to the filter-bank scaling operation to generate warped features, X^{α}, the relation between warped and un-warped features is highly non-linear. Hence Jacobian, $\left|\frac{dX^{\alpha}}{dX}\right|$, is not easy to compute analytically and in practice only the likelihood of the warped features w.r.t. the model, $p(X^{\alpha}|\lambda, U)$, is used for warp-factor estimation after *ignoring the jacobian* (section II for details).

However, the jacobian is a very important term for warp-factor estimation and should not be ignored. In our recently proposed linear transformation (LT) based VTLN, VTLN warping is expressed as a matrix multiplication of un-warped features. Hence in this framework, the jacobian simply turns out as the determinant of the warp matrices and thus gives a framework to compensate jacobian in warp-factor estimation. It has been observed that the jacobian is not a constant term, and hence it should be accounted for.

In this paper, the effect of use of jacobian on Word Recognition Accuracy (WRA) has been evaluated by conducting experiments on different train and test speaker condition on the RM and TIDIGITs task databases. We have used two LT based methods for VTLN, namely L-VTLN and T-VTLN, and have used the jacobian for warp-factor estimation.

In Section II we discuss VTLN as it is done in conventional approach and describe why jacobian is ignored. In Section III, the two approaches for LT implementation of VTLN (L-VTLN and T-VTLN) are discussed. We have also shown how to obtain the jacobian in these methods. Then in Section IV and V, we describe our experimental set up on the RM and TIDIGITs task databases and present our results on compensation of jacobian for warp-factor estimation on different train and test speaker conditions.

II. Vocal Tract Length Normalization (VTLN)

In this section, we describe VTLN for speaker normalization in more detail. Fig.1 shows the effect of variation of vocal tract length on the spectra of a sound when spoken two different speakers. The spectrum shown in solid line is the sound /aa/ spoken by a male speaker and the spectrum shown in dotted line is the same sound when spoken by a female speaker. Clearly, there is significant difference in the two spectra.

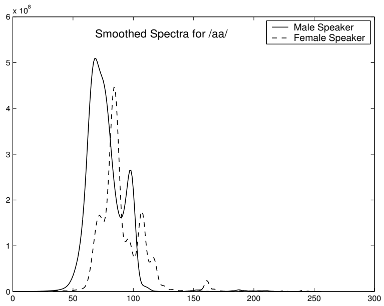


Fig. 1. Spectra of the sound /aa/ when spoken by a male speaker and a female speaker showing variation in the formant frequencies

VTLN tries to reduce this variation in the spectra by warping the spectrum of one speaker so that it approximately matches the other. The frequency-scaling operation in VTLN can be represented as as follows.

$$\tilde{f} = g_\alpha(f) \tag{2}$$

where $g_\alpha(f)$ denotes the frequency-warping applied to the speech spectrum and α is the warping parameter. According to uniform acoustic tube model, the VTLN warping function is usually assumed to be a linear function, i.e.,

$$S_A(f) = S_B(\alpha_{AB} f) \tag{3}$$

In Eq. 3, A and B represent two hypothetical speakers and $S_A(f)$ and $S_B(f)$ their spectrum. Since there is no universal speaker to which the spectra of all speakers can be matched, the VTLN warp-factor, α, is estimated from data in Maximum Likelihood sense. The distribution function of the VTLN warped features can be expressed as,

$$p(X|\lambda^\alpha) = p(X^\alpha|\lambda, U) \quad \left|\frac{dX^\alpha}{dX}\right| \tag{4}$$

where, λ^α, λ, U and $\left|\frac{dX^\alpha}{dX}\right|$ are the VTLN normalized model, SI (or previous iteration VTLN) model, the transcription and the *Jacobian*, respectively. Jacobian comes into picture due to VTLN normalization of the MFCC features. The ML estimate of warp-factor is given by Eq. 1.

In conventional VTLN [1], the frequency-warped MFCC features, X^α, are generated by a filter-bank scaling operation and the maximization over α is done by performing a search over α from 0.8 to 1.20 at the step size of 0.02. Since filter-bank scaling is highly non-linear, Jacobian can not be analytically computed, hence it is *ignored* in practice. Expression for warp-factor estimation after ignoring jacobian is given by

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

Next we discuss the LT based methods for VTLN that allows jacobian to be easily computed.

III. Linear Transformation Approach for VTLN

Recently we have shown that conventional VTLN warping can be expressed as a linear transformation of the un-warped features in several ways. In these methods, the frequency scaling (or, filter-bank scaling operation as in conventional VTLN), can be expressed as a matrix multiplication, i.e.,

$$X^\alpha = g_\alpha(X) \approx W^\alpha X, \tag{6}$$

Following advantages are obtained in LT of VTLN over conventional VTLN:

1) It is not necessary to generate and store the warped features for each values of warp-factor. A VTLN warp-matrix is analytically computed (or estimated from data) for each values of warp-factor. Warped features for a particular value of warp-factor can be generated by linearly transforming the un-warped MFCC features by the VTLN warp matrix corresponding to the warp-factor.
2) Since warping is expressed as linear-transformation, Jacobian can now simply be calculated as the determinant of the VTLN warp-matrix, i.e., $\left|\frac{dX^\alpha}{dX}\right| = |W^\alpha|$

In Linear Transform VTLN frame-work, jacobian can be accounted for in warp-factor estimation, i.e.,

$$\alpha^* = \arg\max_\alpha log\; p(W^\alpha X|\lambda, U) + log(|W^\alpha|) \tag{7}$$

Now, we describe two different methods for implementing linear transformation VTLN that we have used in the experiments in the paper.

A. Linear VTLN (L-VTLN)

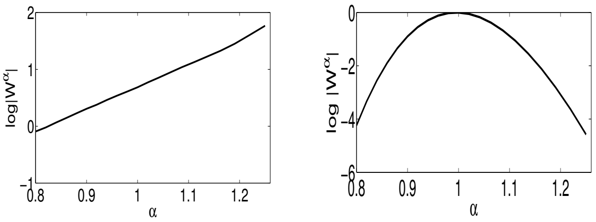
In [2] a method have been proposed to obtain VTLN-warped features, X^α, through a linear transformation of un-warped MFCC features, X. In this method the VTLN matrices are estimated from the training data. The following steps are followed to estimate the matrices:

1) Start with the Non-VTLN model λ_0.
2) The training set is warped in conventional frame-work of VTLN warping with warp-factor α to produce the warped-training set $\{X^\alpha\}_1^R$. R is the number of training utterances.
3) Estimate the (as in CMLLR [3]) transformation, A^α, between the model λ_k and the training set $\{X^\alpha\}_1^R$, i.e.,

$$A^\alpha = \arg\max_A \; p(\{X^\alpha\}_1^R\,|A, \lambda_k) \tag{8}$$

where, the matrix, A, transforms both the means and the co-variances as follows,

$$\hat{\mu}_k = A\; \mu_k\;,\; \hat{\Sigma}_k = A\; \Sigma_k\; A^T \tag{9}$$



(a) L-VTLN (shown for Male Train set of TIDIGITs database)

(b) T-VTLN

Fig. 2. Jacobian of L-VTLN and T-VTLN Warp-matrices

4) Compute the L-VTLN warp-matrix for warp-factor $1/\alpha$ as $W^{\alpha} = (A^{\alpha})^{-1}$.
5) Do warp-factor estimation of the training data following Eq. 7.
6) Train VTLN normalized model, λ_{k+1}, using the warped training utterances.

In L-VTLN, the VTLN warp-matrices, W^{α}, are *estimated* from the training set and the jacobian is simply computed as $|W^{\alpha}|$.

B. Band-limited Interpolation based VTLN (T-VTLN)

In T-VTLN [4], the VTLN warp-matrices are analytically computed, as oppose to being estimated from training data as in the case of L-VTLN. The VTLN warp-matrices, W^{α}, are obtained using the idea of band-limited interpolation, which is given by

$$W^{\alpha} = DT^{\alpha}D^{-1} \tag{10}$$

where D is the DCT transform and T^{α} is the band-limited interpolation matrix, represented by

$$T^{\alpha}_{k,n} = \frac{1}{2N} \sum_{l=0}^{2N-1} e^{-j\frac{2\pi}{2N}\left(\frac{\tilde{\nu_l}}{\nu_s}\right)k} e^{j\frac{2\pi}{2N}\left(\frac{\nu_l}{\nu_s}\right)n}. \tag{11}$$

where ν_l and $\tilde{\nu_l}$ denote the Mel-frequencies corresponding to the physical-frequencies (Hz) before and after frequency scaling, ν_s is the sampling frequency expressed in Mels and N is the number of Mel filters. In this method, the Linear Transformation is exactly computed without any modification in standard MFCC processing.

In T-VTLN, the VTLN warp-matrices, W^{α}, are *analytically computed* using Eq. 10 and the jacobian is calculated as $|W^{\alpha}|$.

In both L-VTLN and T-VTLN methods of VTLN, warp-factor estimation is done by performing a search over α as mentioned in Eq 7. However, it can be done very efficiently in Expectation-Maximization framework [5]. Other advantages of using LT based VTLN can be found in [6], [7].

In Fig. 2(a) and Fig. 2(b) the jacobian for the L-VTLN (estimated using the male training of TIDIGITs database) and T-VTLN warp-matrices (analytically calculated) are shown, respectively. From the plots of the Jacobian, it is clear that jacobian is an important term for warp-factor estimation and it should not be ignored [8].

IV. Results and Discussions

We now present the experimental results on the DARPA Resource Management (RM) and the TIDIGITs databases.

In the case of RM, cross-word tri-phone models were used with decision tree based state tying. The tri-phone HMM models consists of three states, with each state being modeled by 6 diagonal-covariance Gaussian mixture model. A three state model with 6 diagonal-covariance components was used for silence (“sil”), and a single state short-pause (“sp”) model was constructed by tying the state to the center state of the silence model. Training was done using the RM SI-109 training set (for Adult train) that resulted in 1560 states after state tying. For training using Male (or Female), the model was built using the male (or female) part of the RM SI-109 training set. Test was performed on the Feb-89 test set (for Adult test) using RM word pair language model. For test using male (or female), the male (or female) part of the Feb-89 test set was used. 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.

The features in all tasks are 39-dimensional MFCC, 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.

V. Experimental results and Discussions

In this section we present our results of experiments conducted on the RM and the TIDIGITs task databases using L-VTLN and T-VTLN methods and discuss the effect of accounting for jacobian for warp-factor estimation. All VTLN experiments were performed in un-supervised mode, i.e., using the first-pass transcription to compute the likelihood of warped test utterances for warp-factor estimation (unless otherwise mentioned).

A. Experiments using L-VTLN

Different experiments were conducted by taking different train and test speaker conditions using L-VTLN (section III-A). The word recognition accuracy (WRA) are shown in Table I and II for the experiments on RM and TIDIGITs tasks, respectively. In the tables, for example, the experiment M-F were conducted by taking the Male training data to build the acoustic model and tested on Female test data. Similarly, the case A-A denotes Adult Train and Adult test. The results on both the databases are broken into matched and mis-matched train and test speaker conditions. In all cases shown in the table, the L-VTLN matrices were trained using the training data by following the steps described in section III-A.

Analysis of Results on RM Task:
From the experimental results on RM task shown in Table I the following observation are made:

- There is marginal improvement in the matched speaker case of A-A when jacobian was used in warp-factor estimation.

TABLE I

Word Recognition Accuracy (WRA) on RM Tasks for Matched and Mis-matched Train-Test Speaker conditions using L-VTLN

Method	RM Task		
	Matched	Mis-matched	
	A-A	M-F	F-M
No VTLN	96.49	83.43	76.73
L-VTLN (No-Jacob)	97.19	92.82	80.69
L-VTLN (Jacob)	97.23	94.16	83.19

A-A=Adult train - Adult test, M-F= Male train - Female test, F-M=Female train - Male test

TABLE II

Word Recognition Accuracy (WRA) on TIDIGITs Tasks for Matched and Mis-matched Train-Test Speaker conditions using L-VTLN

Method	TIDIGITs Task				
	Matched	Mis-matched			
	A-A	M-C	C-M	M-F	F-M
No VTLN	96.70	68.39	89.12	94.11	94.92
L-VTLN (No-Jacob)	99.64	92.88	94.97	99.36	98.35
L-VTLN (Jacob)	99.62	92.01	96.00	99.20	98.68

A-A=Adult train - Adult test, M-F= Male train - Female test, F-M=Female train - Male test M-C=Male train - Child test, C-M=Child train - Male test

- The performance improvement in the cases of M-F and F-M was significantly high when jacobian was used.

Analysis of Results on TIDIGITs Task:

From the experiments on TIDIGITs task shown in Table II, the following observations are made:

- For the matched case on A-A (Adult Train and Adult Test), use of Jacobian had no effect on the word accuracy.
- There were significant performance improvement in the case of C-M and slight improvement in the case F-M when jacobian was used.
- However there were slight degradation in WRA in the cases of M-C and M-F.

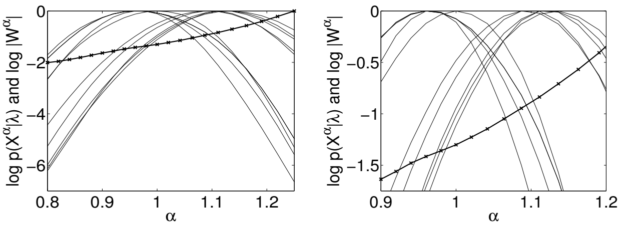
Analysis of Results on A-A:

To understand the reason for jacobian not improving the WRA in the A-A case of TIDIGITs we plotted the log likelihood (without jacobian), log $p(X^\alpha|\lambda)$, for some of the warped test utterances, as a function of warp-factor, α, which are shown in Figure 3. For ease of comparison, the log likelihood scores are normalized so that the maximum is at 0. The (log) jacobian obtained from the L-VTLN matrices, log $|W^\alpha|$, are also shown in the figure after normalization so that the maximum is 0.

From Figure 3(a) we can observe that the variation in likelihood as a function of warp-factor, α, is significantly higher than the variation in jacobian over the entire range of α. Figure 3(b) shows a magnified view of Figure 3(a) shown over the part of the search range of warp-factor where the ML estimate of α is expected to lie. Since the variation of likelihood is very high compared to the variation of jacobian, use of jacobian has no effect and warp-factor estimation is effectively driven by likelihood alone. Therefore, in the A-A case, the WRA when jacobian was used for warp-factor estimation and jacobian not being used were almost identical.

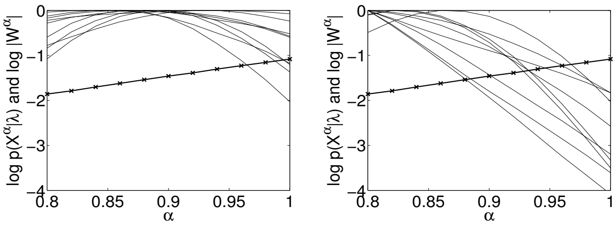
Analysis of Results on M-C and M-F:

In the case of M-C and M-F, there was a degradation in WRA with the use of jacobian. Again, to understand the reason for



(a) Likelihood vs. Jacobian for A-A (b) Magnified View of Fig. 3(a)

Fig. 3. Likelihood (of selected Test utterances) vs. Jacobian of L-VTLN Warp-matrices for Adult Train and Adult Test on TIDIGITs task



(a) First-pass Transcription used for likelihood calculation (b) True Transcription used for likelihood calculation

Fig. 4. Likelihood vs. Jacobian of L-VTLN for M-C when First-pass transcription and True transcription were used during warp-factor estimation of several test utterances in TIDIGITs task

this degradation, the likelihood of warped test utterances and jacobian are shown against warp-factor for M-C in Figure 4.

From Figure 4(a), it can be observed that, unlike the case of A-A, the variation in likelihood is less compared to the variation in jacobian. This is due to the poor first-pass transcription (only 68.39%) that is used for alignment of the test data for warp-factor estimation. Since there are many errors in the first-pass transcription, likelihood of the (children) test data is not computed accurately w.r.t. the (male) model. As a result, the variation in likelihood of the warped test utterances is not large for different values of warp-factor. From the figure it is also observed that the likelihood of the test utterances are dominated by the jacobian function, particularly in the range of α where the ML estimate is expected to lie, and therefore the warp-factor selection is significantly affected by the jacobian causing incorrect estimate of α. This has caused the inferior performance of M-C when jacobian was used.

Effect of use of True Transcription for warp-factor estimation during test in M-C and M-F:

To further check that it is the poor first-pass transcription that is responsible for incorrect estimation of warp-factor, we used the true transcription of the test utterances for warp-factor estimation. The likelihood plots are shown in Figure 4(b) for M-C. The use of true transcription in warp-factor estimation leads to better alignment of the test utterance and hence better computation of likelihood. As it is evident from the figure, the variation in the likelihood increased compared to when first-pass transcription was used.

A degradation in performance in the case of M-F was also observed with the use of jacobian in warp-factor estimation. The WRA for M-C and M-F using true transcription for α estimation are shown in Table III. Now, in both the cases, the

TABLE III

WRA when True Transcription was used for warp-factor estimation for the cases where Jacobian degrades performance in TIDIGITs

Method	Mis-matched	
	M-C	M-F
L-VTLN (No-Jacob)	93.30	99.41
L-VTLN (Jacob)	93.26	99.38

performance with jacobian increased and became very similar to not using jacobian for warp-factor estimation.

B. Experiments using T-VTLN

In Table IV, the experimental results on RM and TIDIGITs Task conducted using T-VTLN (section III-B) are presented.

- In RM task, for the matched case of A-A, use of Jacobian had no effect on the word accuracy.
- There were performance degradation in the case of M-F with the use of jacobian.
- Similarly in TIDIGITs task, jacobian marginally improved the performance in case of A-A.
- There was degradation in WRA in the case of M-C.

TABLE IV

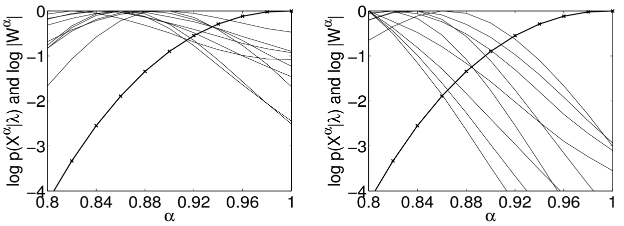
Word Recognition Accuracy (WRA) on RM and TIDIGITs Tasks for Matched and Mis-matched Train-Test Speaker conditions using T-VTLN

	RM task		**TIDIGITs Task**	
	Matched	Mis-Matched	Matched	Mis-matched
	A-A	M-F	A-A	M-C
No VTLN	96.49	83.43	96.70	68.39
T-VTLN (No-Jacob)	97.07	96.65	99.54	96.64
T-VTLN (Jacob)	97.07	96.07	99.58	87.64

Analysis of Results on M-C case of TIDIGITs: Figure 5(a) and Figure 5(b) shows the likelihood of the test utterances for M-C case of TIDIGITs (same set of utterances as used in Figure 4) obtained using the T-VTLN matrices when first-pass transcription and true transcriptions were used for warp-factor estimation, respectively. The jacobian of the T-VTLN matrices are also shown in the figures. From the figure we observe that, due to errors in first-pass transcription, the variation in likelihood of the warped utterances for different values of warp-factor are not very high. As a result, jacobian dominates likelihood and again the effect of likelihood is almost ignored in warp-factor estimation. This caused the degradation in WRA when jacobian is used. Even when the true transcription was used for computation of likelihood of the warped utterances (Figure 5(b)), jacobian had a dominating contribution to warp-factor estimation, causing incorrect α estimation, and hence resulting in inferior performance with jacobian.

VI. Conclusions

In this paper we have studied the effect of use of jacobian in linear transformation based methods of VTLN. By conducting experiments on the RM task and TIDIGITs databases, we observed that when jacobian is used in L-VTLN, the performance improves in all cases of matched and mis-matched



(a) First-pass Transcription used for likelihood calculation

(b) True Transcription used for likelihood calculation

Fig. 5. Likelihood vs. Jacobian of T-VTLN for M-C when First-pass transcription and True transcription were used during warp-factor estimation of several test utterances in TIDIGITs task

speaker conditions of RM task. In TIDIGITs, in L-VTLN cases, use of jacobian improves WRA compared in the cases of F-M and C-M. However, in the cases of M-C and M-F, the degradation in word accuracy is due to errors in first-pass transcription that causes incorrect calculation of likelihood. When true transcription was used for warp-factor estimation in these two cases, WRA performances of using jacobian were similar to that of only likelihood based warp-factor estimation. In T-VTLN based method, on the other hand, in the matched speaker conditions of TIDIGITs and RM task databases (A-A) the performance of use of jacobian was almost similar to only likelihood based warp-factor estimation. In mis-matched cases, however, there was degradation in WRA. We believe that the degradation in performance in T-VTLN with jacobian is due to improper calculations of likelihood, for which jacobian is having dominating contribution to warp-factor estimation. We are looking into this issue in more detail.

VII. 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.

References

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

[2] D. Y. Kim, S. Umesh, M. J. F. Gales, T. Hain, and P. C. Woodland, "Using vtln for broadcast news transcription," in *Interspeech*, September 2004.

[3] 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.

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

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

[6] S. P. Rath, S. Umesh, and A. K. Sarkar, "Using VTLN Matrices for Rapid and Computationally-Efficient Speaker Adaptation with Robustness to First-Pass Transcription Errors," in *Interspeech*, Brighton, September 2009.

[7] S. P. Rath and S. Umesh, "Acoustic Class Specific VTLN-Warping using Regression Class Trees," in *Interspeech*, Brighton, September 2009.

[8] D. R. sanand, S. P. Rath, and S. Umesh, "A Study on the Influence of Covariance Adaptation on Jacobian Compensation in Vocal Tract Length Normalization," in *Interspeech*, Brighton, September 2009.