IDIAP RESEARCH REPORT

STUDY OF JACOBIAN NORMALIZATION FOR VTLN

Lakshmi Saheer Philip N. Garner John Dines

Idiap-RR-25-2010

JULY 2010

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Study of Jacobian Normalization for VTLN

Lakshmi Saheer[1,2], *Philip N. Garner*[1], *John Dines*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne, Switzerland
lsaheer@idiap.ch, pgarner@idiap.ch, dines@idiap.ch

Abstract

The divergence of the theory and practice of vocal tract length normalization (VTLN) is addressed, with particular emphasis on the role of the Jacobian determinant. VTLN is placed in a Bayesian setting, which brings in the concept of a prior on the warping factor. The form of the prior, together with acoustic scaling and numerical conditioning are then discussed and evaluated. It is concluded that the Jacobian determinant is important in VTLN, especially for the high dimensional features used in HMM based speech synthesis, and difficulties normally associated with the Jacobian determinant can be attributed to prior and scaling.
Index Terms: Jacobian Normalization, Vocal Tract Length Normalization, Adaptation

1. Introduction

Automatic speech recognition (ASR) has long been based on hidden Markov models (HMMs). HMMs for ASR are generally combined with some kind of feature or model adaptation techniques. The effect of the adaptation is to move the characteristics of the observed features (the speech to be recognised) closer to those of the model. Generally the model is speaker independent. In an ideal case, under speaker adaptive training (SAT), the model represents an average voice (canonical model) and the adaptation represents a transformation from the average (characterless) voice to the voice of a given person. By contrast, speech synthesis (often referred to as text to speech, or TTS) has tended to rely on unit selection techniques. Such techniques involve joining together small chunks of the speech of a given person to form an utterance. Unit selection has no concept of speaker independence; there is one speaker only.

Recently, HMMs have been shown to be capable of performing TTS too, and with care can produce synthetic speech of a quality comparable to unit selection. This in turn brings the possibilities of adaptation to TTS. A stored average voice can be transformed to sound like a voice represented by a given transform. Such transforms are typically linear transforms produced by maximum likelihood linear regression (MLLR), or one of its derivatives. The advantage of linear transforms over other possibilities such as maximum a-posteriori (MAP) model adaptation is that they require much less adaptation data; of the order of a few minutes.

Vocal tract length normalization (VTLN) is another adaptation technique. It is based on the physical observation that vocal tract length varies across speakers from around 18cm in males to around 13cm in females. Formant frequencies are inversely proportional to vocal tract length, and hence can vary by around 25%. Although implementation details differ, VTLN is generally characterised by a single parameter that warps the spectrum towards that of an average vocal tract in much the same way that MLLR transforms can warp towards an average voice. The parameter has been shown experimentally to have a bimodal distribution with the modes representing male and female speech. Crucially for this study, VTLN has been shown to be a linear transform in the feature (cepstral) domain [1, 2, 3]. VTLN, with just one degree of freedom, hence represents a linear transform that can adapt with very few adaptation data; of the order of a few seconds, or a single utterance.

Being a feature transformation, training VTLN in a statistical sense requires calculation of a Jacobian determinant. Whilst this can be proved easily and is acknowledged in the literature, many authors either cannot consider it due to the nature of the transform, or choose not to consider it. Typically, the effect of the Jacobian determinant is either minimal or can be removed by cepstral mean and variance normalization (CMVN) [4, 5, 6]. When using ASR-like adaptation in HMM based TTS, CMVN is generally not used because it represents a distortion of the speech characteristics. Further, the high feature dimensionality associated with TTS leads to more extreme values of the Jacobian determinant, meaning in turn that it cannot be ignored.

In the following sections, VTLN is formulated in a Bayesian sense. Some reasons for the divergence of theory and practice in the literature are discussed, and possible solutions considered. Results are presented as histograms of warping

factors, with ASR results confirming a correlation with the histograms.

2. VTLN Formulation

VTLN is characterised by

- A warping function, e.g., linear, piecewise linear, non-linear, bilinear.
- A warping factor, typically denoted α.
- An optimization criterion, e.g., MAP, ML.

One of the main advantages of VTLN is that the warping factor can be reliably estimated even with a single adaptation sentence for each test speaker.

A brute force way of computing the warping factor for each speaker is the ML based grid search technique [4]:

$$\hat{\alpha}_{s1} = \underset{\alpha}{\operatorname{argmax}}\, p\left(\boldsymbol{x}_{\alpha_{s1}} \mid \boldsymbol{\Theta}, \boldsymbol{w}_{s1}\right) \tag{1}$$

where $\boldsymbol{x}_{\alpha_{s1}}$ represents the features warped with the warping factor α_{s1}, which is the warping factor for speaker s1. $\boldsymbol{\Theta}$ represents the model and $\boldsymbol{w}_{s1}$ represents the transcription corresponding to the data from which the features are extracted for speaker s1. $\hat{\alpha}_{s1}$ represents the best warping factor for the same speaker.

In the work of Pitz and colleagues [2, 3], it is argued persuasively that VTLN amounts to a linear transform in the cepstral domain. In fact, this *bilinear transform* based warping function is also evident from the mel-generalised approach to feature extraction [7], that is useful in speech synthesis. It has also been studied by Uebel and Woodland [6] and by McDonough [8]. The matrix formulation for the bilinear function can be derived from the recursions of the Mel-Generalized Cepstral features and is of the form:

$$\boldsymbol{A} = \begin{bmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{M_1} \\ 0 & 1-\alpha^2 & 2\alpha(1-\alpha^2) & \cdots & M\alpha^{M_1}(1-\alpha^2) \\ 0 & -\alpha(1-\alpha^2) & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & (-1)^{N_1}(1-\alpha^2)\alpha^{N_1} & \cdots & \cdots & \cdots \end{bmatrix}$$

where $M_1 = M - 1$ and $N_1 = M - 1$.

The matrix formulation enables the calculation of Jacobian normalization as the determinant of the transformation matrix.

2.1. Probabilistic formulation

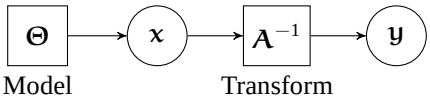


Figure 1: Generative model for vocal tract "warping"

Assume a model, $\boldsymbol{\Theta}$, generates a sample, $\boldsymbol{x}$. The sample is then distorted by a linear transform, $\boldsymbol{A}^{-1}$, a function of α, to give an observation $\boldsymbol{y} = \boldsymbol{A}^{-1}\boldsymbol{x}$. Here, we follow convention where $\boldsymbol{A}$ is a feature transform so the generative transform is $\boldsymbol{A}^{-1}$. The goal is to find an optimal value, $\hat{\alpha}$, of α. Bayes's theorem gives the maximum *a posteriori* estimator:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}}\, p\left(\alpha \mid \boldsymbol{y}, \boldsymbol{\Theta}\right) \propto p\left(\boldsymbol{y} \mid \alpha, \boldsymbol{\Theta}\right) p\left(\alpha \mid \boldsymbol{\Theta}\right). \tag{2}$$

To evaluate the first term on the RHS of equation 2, notice that the model generates $\boldsymbol{x}$ rather than $\boldsymbol{y}$, so it needs a change of variable $\boldsymbol{y} \rightarrow \boldsymbol{x}$. The Jacobian determinant for the change of variable is,

$$J = |\boldsymbol{A}|, \tag{3}$$

where the notation is taken to mean the determinant of the matrix. So,

$$p\left(\boldsymbol{y} \mid \alpha, \boldsymbol{\Theta}\right) = |\boldsymbol{A}|\, p\left(\boldsymbol{A}\boldsymbol{y} \mid \alpha, \boldsymbol{\Theta}\right). \tag{4}$$

The second term on the RHS of equation 2 is a prior on α. Notice that α is actually independent of the model, $\mathbf{\Theta}$, so it could be written unconditional. However, α is posterior to the training data, $\mathbf{D}$, that was used to train $\mathbf{\Theta}$. So, equation 2 can be evaluated as

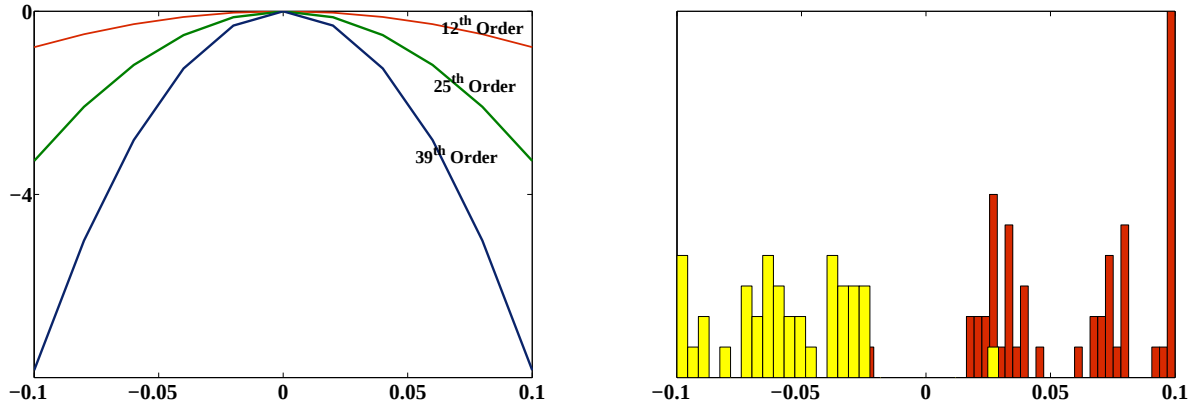
$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}} |\boldsymbol{A}| \, \mathrm{p}\left(\mathbf{A}\mathbf{y} \mid \alpha, \mathbf{\Theta}\right) \mathrm{p}\left(\alpha \mid \mathbf{D}\right). \tag{5}$$

Notice that the "prior" is not normally considered. It will be discussed later.

2.2. Issues with Jacobian Normalization

Jacobian normalization usually tends to degrade the warping factor estimation and thus has adverse effects on the performance of VTLN. It can be seen in the literature that most VTLN implementations either ignore the Jacobian normalization or replace it with cepstral mean/variance normalization [6, 4, 9]. One recent study [10] on mismatched train and test conditions addressed this issue by compensating with a variance adaptation on top of VTLN. This approach was a conclusion of the fact that the VTLN transformation on both mean and variance does not fully match the data when there is a mismatch in speaker conditions which resulted in the degradation of performance when using Jacobian normalization. Similar problems are observed when applying VTLN on higher order features (of the order of 25 or 39) which are used for statistical speech synthesis [11]. The effect is less noticeable for lower order features (of the order of 12 or 15), which are usually used in ASR. The effect of Jacobian normalization for different feature orders can be visualized in Figure 2a. The Jacobian is used as $\log |\boldsymbol{A}|$. It can be seen that the values have a flatter distribution for lower order features and variations become more prominent for higher order features. The effect of not using Jacobian normalization for 39^{th} order features is shown in Figure 2b. As opposed to Figure 3a, the warping factors tend to spread towards boundary.

When using grid search based warping factor estimation, features with different warping factors are aligned with the transcription for estimating the likelihood scores. Usually, a standard speech toolkit like HTK is used to generate the alignments. This results in changes in alignment boundaries for different warping factors. It was noted that there was change of about 1 frame in most of the boundaries, where the frame shift was 5ms. In experiments to estimate the warping factors with a fixed alignment, it was found that the alignment changes do not have a significant impact on the warping factor estimation. The alignment problem does not exist when warping factor estimation is embedded in the EM training algorithm because it performs calculations on the statistics generated from a fixed alignment.



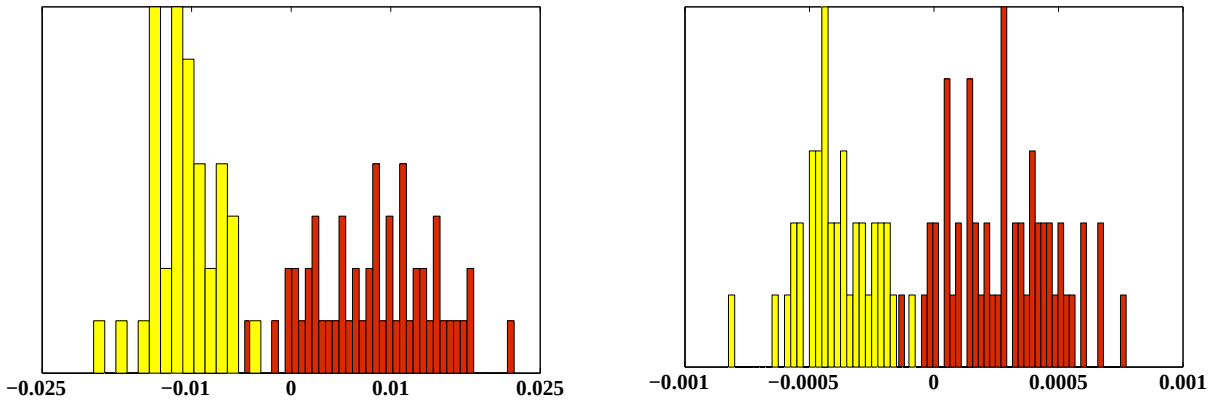
(a) Jacobian value

(b) No Jacobian normalization

Figure 2: Jacobian value calculated as $\log |\boldsymbol{A}|$ for various feature dimensions. Distributions over warping factor value without Jacobian normalization. The abscissa is α, the warping factor.

3. Experimental

The following experiments used the WSJCAM0 British English read speech database. 39^{th} order acoustic models for TTS were trained using hidden semi-Markov models (HSMM) with only a single Gaussian PDF per state. The bilinear transform based VTLN was applied on the mel-cepstral (MCEP) features with a warping factor within the range of -0.1 to 0.1. In all cases, VTLN warping was estimated using the EM approach similar to [5].



(a) Full matrix determinant

(b) Diagonal determinant

Figure 3: Warping factors estimated from 39^{th} order features with different Jacobian normalizations. The abscissa in all cases is α, the warping factor, although note that the ranges vary. The ordinate depends on the plot; see the text for details.

As the theory suggests, Jacobian normalization should be used for warping factor estimation. When the feature stream contains dynamic components, the transformation can be expressed as follows.

$$\hat{c} = \begin{bmatrix} \boldsymbol{A} & 0 & 0 \\ 0 & \boldsymbol{A} & 0 \\ 0 & 0 & \boldsymbol{A} \end{bmatrix} \begin{bmatrix} c \\ \Delta c \\ \Delta^2 c \end{bmatrix}, \tag{6}$$

where $\boldsymbol{A}$ is the transformation on the static features and can be directly applied to the dynamic part of the cepstra as well. It can be shown experimentally that the warping factors estimated from the static features are more accurate. Estimating warping factors as a transformation on the cepstrum should take into account the fact that the feature stream usually contains dynamic features which can disrupt the warping factor estimation. Table 1 shows the warping factors estimated using static and dynamic feature vectors separately for a male and female speaker. This problem should not be observed

Gender	**Static**	Δ	Δ^2
Male	0.0195	0.0100	-0.0145
Female	-0.0260	-0.0142	0.0134

Table 1: Warping factors for components of feature vectors

in VTLN techniques embedded into the feature extraction step (like warping the filter banks of MFCC features) which estimate dynamic features from the warped static features.

3.1. Probable causes of degradation

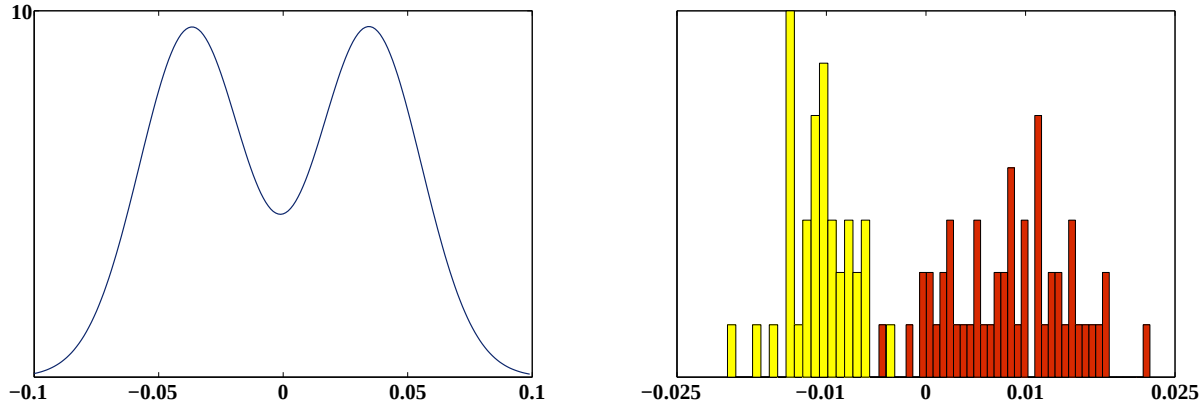
In the following sections, some issues regarding application and calculation of Jacobian determinant are discussed.

3.1.1. Erroneous use of flat prior

The prior on the warping factor in equation 5 has previously been ignored. Ignoring a prior normally corresponds to assuming a flat prior. Where many data are available, this is often a reasonable approach. Conversely, where few data are available, priors can be important. It is simple to argue subjectively that the prior on the warping factor should not be flat:

- It should tend to zero at the extreme values ± 1.
- It should be bimodal, representing male and female speech.

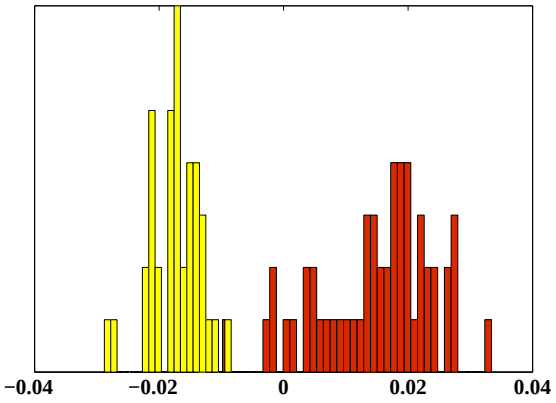
Objectively, the prior can be measured via a histogram of warping factors calculated over a large number of speakers, for each of whom a large amount of data exists. Such histograms are shown in Figure 3a, and moments can be measured to



(a) Beta mixture

(b) Bimodal prior

Figure 4: Beta mixture prior on α. Distributions over warping factor value with beta prior. The abscissa in all cases is α, the warping factor, although note that the ranges vary. The ordinate depends on the plot; see the text for details.



(a) Scale factor

Figure 5: Warping factors estimated from 39^{th} order features with a scale factor of 2 for the likelihoods. The abscissa is α, the warping factor.

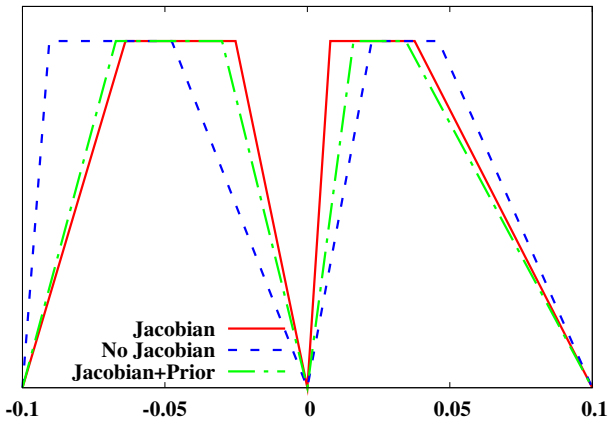
infer a parametric distribution. Here, we use a two-component beta mixture, transformed to span the range ± 1:

$$\mathrm{p}\left(\alpha \mid \mathbf{D}\right) \propto \sum_{\mathrm{g} \in \{\mathrm{m},\mathrm{f}\}} (1+\alpha)^{\mathrm{p_g}-1}(1-\alpha)^{\mathrm{q_g}-1}, \tag{7}$$

where $\{\mathrm{p_m}, \mathrm{q_m}\}$ and $\{\mathrm{p_f}, \mathrm{q_f}\}$ are the pairs of beta parameters for male and female speech respectively, as in Figure 4a.

Notice that *not* using a Jacobian determinant has the effect of using a prior with a PDF proportional to the inverse of the Jacobian (c.f. Figure 2a). This is, in some sense, a bimodal distribution. Certainly it biases α away from zero, enhancing the relative separation of the male and female modes.

It can be seen from Figure 4b that the prior does not have much impact on warping factor estimation for the training data. The changes are expected to be seen only in the warping factors for the test data. This in turn could explain why it has been observed by earlier researchers that not using Jacobian normalization improves performance especially in testing. During testing, the data is insufficient to generate a meaningful likelihood.



(a) Test data

Figure 6: Warping factors estimated for test data from 13^{th} order features with and without Jacobian normalization and prior. The abscissa is α, the warping factor.

3.1.2. Underestimation of acoustic likelihood

In Large Vocabulary ASR, it is common to use a language model match factor that in fact compensates for the acoustic likelihoods being too small. This in turn is because successive acoustic frames have much more correlation than the HMM can model. Applied more correctly to the acoustic calculation, we might expect that the correction should apply to the likelihoods but not to the Jacobian. In fact, this was investigated by Pitz [1], who applied the factor to the Jacobian analogous to the language model scale. This suggests an estimator of the form

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}} \left[\sum (\mathrm{LL}) \times \mathrm{sf} + \mathrm{F} \times \log |\mathbf{A}| \right], \tag{8}$$

where, LL represents the log-likelihood score, sf represents the scale factor for boosting likelihoods, F represents the total number of frames, $\mathbf{A}$ represents the transformation matrix and α represents the warping factor. The effect of the scale factor value '2' is shown in Figure 5a. There is no standard formula for calculating the scale factor; it is estimated empirically.

3.1.3. Numerical conditioning

The matrix formulated for expressing VTLN as a linear transformation of the cepstrum can be "large", especially when using higher order cepstral features. As the order of the cepstra and the value of the warping factor increases, the calculation of the determinant becomes numerically unstable. The log of the determinant can be calculated as accurately as the sum of the logs of the eigen values of the matrix thus:

$$\log |\mathbf{A}| = \frac{1}{2} \sum_{i=1}^{N} \log (e_i \times e_i^*) \tag{9}$$

where, e_i represents an eigen value of the matrix $\mathbf{A}$ and e_i^* represents the conjugate of the complex number e_i.

Alternatively, it can be observed that the matrix is diagonally dominant. Hence, a diagonal covariance assumption could be used, discarding the non-diagonal elements. The resulting determinant can be estimated as the product of the diagonal elements of the matrix. Instead of ignoring non-diagonal elements, terms with higher order powers of warping factor could also be ignored [12, 13]. These work arounds give a closed form solution to the auxiliary function when VTLN is formulated as an EM optimization. This reduces the time complexity of the warping factor estimation. Both these cases increase the effect of the Jacobian and give warping factor values a push in the wrong direction. This can be observed from Figure 3b, where the warping factors of male and female speakers are pushed towards a region of minimum warping. It also results in no proper distinction between the warping factors of the two genders.

3.2. Recognition performance

Speech recognition experiments are presented here showing that Jacobian normalization should be used in VTLN. The hidden Markov models were built with 13 dimensional cepstral features with Δ and Δ^2 for the (US English) WSJ0 database. The models were built using single mixture PDFs to demonstrate the maximum impact of VTLN, and because only single mixture models can be used in synthesis. In particular, this was to avoid the situation where multi-mixture models either over-fitted, or modeled the speaker variabilities that could be attributed to VTLN. The performance differences are not statistically significant, but support the fact that Jacobian normalization should not degrade the performance. These results are similar to ones shown in [1], where experiments are performed using a scale factor for the Jacobian analogous to inverse scaling of the likelihood.

Moments of symmetric beta prior distributions were estimated from the warping factors for the conversational speech database presented in [9]. The warping factors for the test speakers are shown in Figure 6a, which shows that using a prior distribution can estimate similar warping factors to those that might result from not using Jacobian normalization. The recognition performance is not significantly affected by the separation of the warping factors. However, the separation is important in statistical speech synthesis, which demands distinct warping factors for each speaker to bring as many characteristics of the speaker as possible in the synthesized speech. A scaling factor may be needed in TTS where the prior is insignificant when combined with the large likelihood scores produced by the higher order features.

SI-model	**VTLN**		
	No Jacobian	**Jacobian**	**Jacobian+Prior**
22.16	19.43	19.33	19.49

Table 2: WER for 13^{th} order features on the Nov93 Eval

4. Conclusions

VTLN is a powerful speaker normalization technique that can give performance improvements in both statistical speech recognition and synthesis. This paper emphasizes the fact that Jacobian normalization is an important component of VTLN and should not be ignored, especially for higher order features. Different possible causes for the degradation of warping factor estimation when using Jacobian normalization were investigated. It was shown that not using the Jacobian has a similar effect to using a strong prior for the warping factors. Thus, not using a prior creates more problems in testing where the amount of data is insufficient to completely mask the effect of the prior. The right usage of Jacobian normalization combined with including a proper prior can give similar warping factor distributions as not using a Jacobian. Though the results are shown with bilinear transforms, the same should hold for other warping functions as well.

5. Acknowledgements

The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project).

6. References

[1] M. Pitz, "Investigations on linear transformations for speaker adaptation and normalization," Ph.D. dissertation, AACHEN University, 2005.

[2] M. Pitz, S. Molau, R. Schlüter, and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," in *Proceedings of EUROSPEECH*, 2001, pp. 2653–2656.

[3] H. Ney, U. Essen, and R. Kneser, "On the estimation of 'small' probabilities by leaving-one-out," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 12, pp. 1202–1212, December 1995.

[4] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proceedings of ICASSP*, Washington, DC, USA, 1996, pp. 353–356.

[5] S. Panchapagesan and A. Alwan, "Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC," *Computer Speech & Language*, vol. 23, no. 1, pp. 42–64, 2009.

[6] L. F. Uebel and P. C. Woodland, “An investigation into vocal tract length normalisation,” in *Proceedings of the European Conference on Speech Communication and Technology*, Budapest, Hungary, 1999, pp. 2527–2530.

[7] K. Tokuda, “Mel-generalized cepstral analysis —a unified approach to speech spectral estimation,” in *Proceedings of International Conference on Spoken Language Processing*, Yokohama, Japan, 1994, pp. 1043–1046.

[8] J. W. McDonough, “Speaker compensation with all-pass transforms,” Ph.D. dissertation, John Hopkins University, 2000.

[9] G. Garau, “Speaker normalization for large vocabulary multiparty conversational speech recognition,” Ph.D. dissertation, University of Edinburgh, 2008.

[10] 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 *Proceedings of Interspeech*, Brighton, UK, 2009, pp. 584–587.

[11] L. Saheer, P. N. Garner, J. Dines, and H. Liang, “VTLN adaptation for statistical speech synthesis,” in *Proceedings of ICASSP*, Dallas, Texas, USA, 2010, pp. 4838–4841.

[12] T. Emori and K. Shinoda, “Rapid vocal tract length normalization using maximum likelihood estimation,” in *Proceedings of Eurospeech*, 2001, pp. 1649–1652.

[13] M. Hirohata, T. Masuko, and T. Kobayashi, “A study on average voice model training using vocal tract length normalization,” *IEICE Technical Report,*, vol. 103 (27), pp. 69–74, 2003, in Japanese.