Implementation of VTLN for Statistical Speech Synthesis

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

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

Abstract

Vocal tract length normalization is an important feature normalization technique that can be used to perform speaker adaptation when very little adaptation data is available. It was shown earlier that VTLN can be applied to statistical speech synthesis and was shown to give additive improvements to CMLLR. This paper presents an EM optimization for estimating more accurate warping factors. The EM formulation helps to embed the feature normalization in the HMM training. This helps in estimating the warping factors more efficiently and enables the use of multiple (appropriate) warping factors for different state clusters of the same speaker.

Index Terms: Vocal tract length normalization, Expectation Maximization Optimization, HMM Synthesis, Adaptation

1. Introduction

Hidden Markov model (HMM) is a popular technique used in automatic speech recognition (ASR). Speaker independent (SI) models are built by estimating the parameters of HMM using data collected from a large number of speakers. Model adaptation techniques entail linear transformation of the means and variances of an HMM to match the characteristics of the speech for a given speaker. The same techniques can be used to remove the inter-speaker variability in the training data. The resulting speaker adaptive (SAT) models have better performance than the SI models in ASR. Feature adaptation, on the other hand, transforms the feature vectors rather than the model parameters. The effects of model adaptation can be accomplished to some extent using feature adaptation techniques (also widely known as speaker normalization techniques). The main advantage of speaker normalization is that the number of parameters to be estimated from the adaptation data is generally smaller compared to the standard model based adaptation techniques. Hence, adaptation can be carried out with very little adaptation data.

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 [1]. A stored average voice can be transformed to sound like a voice represented by the transform for a given speaker. Such transforms are typically linear transforms similar to the ones used in ASR. Speaker normalization techniques can also be used in TTS to generate adapted speech using very little adaptation data; of the order of a few minutes.

Vocal tract length normalization (VTLN) is inspired from the physical observation that the vocal tract length (VTL) varies across different speakers in the range of around 18 cm in males to around 13 cm in females. The formant frequency positions are inversely proportional to VTL, and hence can vary around 25%. Although implementation details differ, VTLN is generally characterized by a single parameter that warps the spectrum towards that of an average vocal tract in much the same way that maximum likelihood linear regression (MLLR) transforms can warp towards an average voice.

An efficient implementation of VTLN using expectation maximization (EM) with Brent's search optimization for synthesis is presented in this paper. Optimal warping factors for synthesis are analyzed, and techniques to estimate similar warping factors from the model are examined. Problems with Jacobian normalization for VTLN warping factor estimation are briefly discussed along with a technique that achieves best performance for synthesis. This paper also investigates the multi-class EM-VTLN estimation in the context of statistical synthesis. The features used for statistical speech synthesis have very high dimensionality (of the order of 25 or 39) when compared to ASR features. There are some issues with VTLN estimation for higher order features which were presented in earlier work [2] and further investigated here.

2. VTLN

The main components involved in VTLN are: a warping function, a warping factor and an optimization criterion. The all-pass transform approximates most commonly used transformations in VTLN [3, 4]. The bilinear transform based warping function has only a single variable α as the warping factor which is representative of the ratio of the VTL of the speaker to the average VTL. The terms warping factor and 'α' refer to the same parameter and are used interchangeably throughout this paper. A brute force way of computing the warping factor for each speaker is the maximum likelihood (ML) based grid search technique. ML optimization is given by [5]:

$$\hat{\alpha}_s = \arg\max_{\alpha} p(x_{\alpha_s} \mid \Theta, w_s) p(\alpha \mid \Theta) \tag{1}$$

where x_{α_s} represents the features warped with the warping factor α_s, which is the warping factor for speaker s. Θ represents the model and w_s represents the transcription corresponding to the data from which the features are extracted for speaker s. $\hat{\alpha}_s$ represents the best warping factor for the same speaker. $p(\alpha|\Theta)$ is the prior probability of α for a given model.

Preliminary results using VTLN in statistical speech synthesis are presented in [2]. The bilinear transform based warping function is used in an ML optimization framework using a grid search technique. The all-pass transform based normalization is applied to the mel-generalized cepstral (MGCEP) features that are commonly used in statistical speech synthesis. It is shown that VTLN brings in some speaker characteristics and provides additive improvements to CMLLR, especially when there is a limited number of adaptation utterances. In [6], it is argued persuasively that VTLN amounts to a linear transform in the cepstral domain. In fact, this is also evident from the mel-generalized approach to feature extraction [7]. Hence, VTLN can also be implemented as an equivalent model transform. Representation of VTLN as a model transformation enables the use of techniques like EM for finding the optimal warping factors [8, 9]. The main advantage of using EM is that the resulting warping factor estimation is based on a gradient descent technique which provides finer granularity of α values. This implementation is also efficient in time and space, since features need not be recomputed for every warping factor. EM can be embedded into the HMM training utilizing the same sufficient statistics as CMLLR. This also opens up the possibility of estimating multiple warping factors for different phone classes. Since the ML optimization does not provide a closed form solution to the EM auxiliary function, Brent's search is used to estimate the optimal warping factors.

The Jacobian determinant should be used in the likelihood calculation to choose a warping parameter, α; this is attributed to [10]. Representation of VTLN as a linear transform facilitates a simple estimation of this factor. In fact, [10] only use the Jacobian as part of a more involved derivation of an algorithm to train the transformation. TTS uses higher order features compared to ASR which results in some challenges in warping factor estimation [2]. Jacobian normalization also causes some problems in warping factor estimation. A detailed study of techniques to overcome the problems with Jacobian normalization is presented in [11] along with a Bayesian interpretation of VTLN. Jacobian normalization and higher order features together reduce the spread of α values and limit them to a range corresponding to negligible warping.

It was shown in [11] that techniques to improve the spread of warping factors did not show significant improvements in performance of ASR. Unlike ASR, it is not easy to decide which warping factors can give better performance for TTS. Ideally, TTS should favour higher values of αs since this brings in strong gender characteristics. A few informal perceptual experiments are conducted at the begining of this research which resulted in the following observations. It is observed that male or female characteristics could be perceived only when the warping is beyond a certain limit. Also, it is hard to discriminate the speakers with similar VTLs. Noticeable changes in the characteristics of the synthesized voice are observed only if the warping factors have a higher interval. In TTS, VTLN alone cannot bring in many speaker characteristics. There might even be a correlation between pitch and warping factor estimation that needs to be explored. Towards this end, this paper first presents the results of a subjective evaluation designed to assist in finding optimal warping factors for VTLN adaptation of HMM-based TTS.

3. Analysis of VTLN for synthesis

VTL varies across speakers resulting in corresponding changes in the spectral peak positions. Alternatively, warping the spectral frequencies should bring in approximately the same variation that is audible due to the differences in VTL. A preliminary experiment conducted on a speaker's voice using analysis synthesis with different levels of warping provides evidence for this fact. It was noticed that whenever the spectral frequencies are expanded, the speech sounded more "feminine" as if from a shorter vocal tract. Also, whenever the spectral frequencies are compressed, the speech sounded more "masculine" as if from a longer vocal tract. Both phenomena are observed in spite of using the natural pitch of the speaker. These observations led to the design of a subjective evaluation to determine the optimal warping factors for a set of speakers. The values obtained from these evaluations are compared with the warping factors derived from the model.

Table 1: Frequency of female speakers with different combinations of vocal tract length and pitch

Pitch vs. Alpha group	**1**	**2**	**3**	**4**	**5**	**6**
Low (159-190)	1	5	4	1	0	0
Medium (191-222)	0	4	8	8	1	1
High (223-255)	0	1	1	4	0	1

3.1. Experimental design

The HMM speech synthesis system (HTS) [12] is used to build average voice models using 39^{th} order cepstral features along with Δ and Δ^2 values of MGCEP features. Experiments are performed on the WSJCAM0 (British English) database with 92 speakers in the training set. The details of the synthesis system can be seen in [2].

A set of 20 speakers are selected from the 40 female speakers present in training in such a way that they covered the different possible combinations of pitch and VTLs. The gender restriction helps to minimize the size of the evaluations. The distribution of the warping factors for male speakers is expected to be symmetric to that of the female speakers. The pitch range of all the female speakers in the training data is equally divided into 3 sets: high, medium and low. Similarly, the range of α values derived using the SI model for these speakers is also divided into 6 equally spaced groups. The warping factors are estimated using the EM approach described in the next section. Jacobian normalization is used for estimating the α values from the average voice HMM models. The distribution of speakers according to this grouping is shown in Table 1. 20 speakers are selected using this table so that maximum possible combinations are covered. It can be observed that the frequency of speakers with high pitch and low warping is small compared to the combination of high pitch and high warping. This suggests the idea that mostly high pitch voices are associated with females who have shorter VTL compared to males.

Natural pitch contours are extracted from the recorded speech of the selected speakers. Speech files are synthesized using the average voice models and the original pitch contours with 6 different warping factors in the range 0 to 0.1. Listeners are asked to judge the speaker similarity in the original speech file with that of the speech synthesized with different warping factors and the natural pitch of the speaker. This is repeated for 20 utterances each from a different speaker. It is interesting to note that combination of a single pitch with different vocal tract lengths can generate a wide variety of voices. A few expert listeners could perceive that the speaker's voice can be almost reproduced from the average voice with the natural pitch and just a single parameter representing the VTL.

3.2. Results and Discussion

25 listeners participated in this evaluation. The results of the subjective evaluations for 20 speakers are shown in Figure 1a. Each box represents a speaker in the evaluation set. Listeners prefer higher warping factors rather than lower warping factors. The extreme warping is also not preferred. Correlation between results from subjective evaluations and α derived from the HMM models are presented in Table 2. The table compares the values of mean, mode and median of the warping factors

Table 2: Correlation between model derived αs (with and without Jacobian) and results of subjective evaluation. Correlation between warping factors from both schemes and pitch is also presented.

	Pitch	**Mean**	**Mode**	**Median**
Jacobian	-0.4875	0.2238	0.0553	0.2154
No Jacobian	-0.3396	0.4362	0.1976	0.4821
Pitch	-	-0.1244	-0.1120	-0.0400

observed in the subjective evaluation. The results are not statistically significant as it shows there is no significant correlation between any values. The best correlation is seen between the means of the warping factors from subjective evaluation and those derived from the model without using Jacobian normalization. A more detailed study of how Jacobian normalization affects the warping factor estimation is presented in section 5.2. Pitch does not show much correlation to warping factors derived using any scheme. The model derived warping factors have closer correlation to pitch than the warping factors derived from the subjective evaluations. Further investigation is required on this subject before making any further conclusions regarding the relation between pitch and VTL. Perception of VTLN is a very difficult task to assess, but it is evident that on average the perceived warping factors are higher than those estimated with the model using ASR paradigm. This experiment provides a good prior distribution of the warping factors for VTLN synthesis. The next section presents the details of an efficient VTLN implementation for HMM synthesis using the EM optimization.

4. EM based warping factor estimation

It has been shown that EM can be used to estimate VTLN warping factors for ASR [3, 8, 9]. Warping parameters are estimated by maximizing the EM auxiliary function over the adaptation data. The objective function obtained is similar to the one used in MLLR or CMLLR [13]. Even the same sufficient statistics as used in CMLLR can be used for optimizing the VTLN auxiliary function.

Earlier research applying a grid search based bilinear transform VTLN using ML criteria for statistical speech synthesis is presented in [2]. There are many drawbacks in the grid search approach for warping factor estimation, the first one being that warping factors are chosen from a set of available values and cannot be estimated with greater precision. Another drawback is that the likelihood estimation for features with different warping factors consumes a lot of processing time and requires features to be extracted for each warping factor in the grid. This increases the complexity of training process in time and resources. This work presents an EM formulation for the warping factor estimation which performs a gradient descent rather than grid search. This enables more accurate estimation of warping factors and embeds this estimation in the HMM training. The EM formulation exploits the representation of VTLN as a model transform and does not involve calculation of features with different warping factors. Hence, the warping factors can be estimated very efficiently and accurately.

4.1. VTLN as model transform

It was shown in [6] that VTLN can be represented as a linear transform of the cepstral features. Warping the spectral frequencies can be represented equivalently as a feature transform.

$$x_\alpha = A_\alpha \times x \tag{2}$$

where, x_α are spectral features x warped with the warping factor α which can be represented as a matrix transform denoted by A_α. This is equivalent to

$$c_\alpha = \begin{bmatrix} A_\alpha & 0 & 0 \\ 0 & A_\alpha & 0 \\ 0 & 0 & A_\alpha \end{bmatrix} \begin{bmatrix} c \\ \Delta c \\ \Delta^2 c \end{bmatrix} \tag{3}$$

where, c_α is the warped cepstral coefficients, c is the static features, Δc and $\Delta^2 c$ are dynamic part of the cepstra. Transformation can be directly applied to the dynamic part of the cepstra as well. The unwarped cepstral features are multiplied with the linear transformation matrix to generate warped features. This results in significant computational savings since features need not be individually recomputed for each warping factor. Matrix representation of the MGCEP bilinear transform in cepstral domain was presented in [2].

Similar to the CMLLR adaptation, feature transform can be analogously represented as a model transform [13]. The maximum likelihood optimization in feature domain is:

$$\hat{\alpha} = \arg\max_{\alpha} p(A_\alpha x|\mu, \Sigma) p(\alpha|\Theta) \tag{4}$$

The same equation can be represented as a model transform:

$$\hat{\alpha} = \arg\max_{\alpha} |A_\alpha| p(x|A_\alpha^{-1}\mu, (A_\alpha^{-1})^T \Sigma A_\alpha^{-1}) p(\alpha|\Theta) \tag{5}$$

μ and Σ correspond to the mean and variance of a gaussian component in the model. The Jacobian normalization can be calculated as the determinant of matrix (A_α) representing the linear transformation of the cepstral features.

4.2. Auxiliary function for EM

The EM formulation of warping factor estimation results in the following auxiliary function. Taking the log of the function and considering the assumption of Gaussian components in the model.

$$\hat{\alpha} = \arg\max_{\alpha} \left\{ \sum_{f=1}^{F} \sum_{m=1}^{M} \gamma_m \Big[\log(N(A_\alpha x|\mu_m, \Sigma_m)) \right.$$
$$\left. + \log|A_\alpha| \Big] + \log p(\alpha|\Theta) \right\} \tag{6}$$

where, A_α is the transformation matrix for input feature vector x, M is the total number of mixtures, F is the total number of frames, γ_m is the posterior probability of mixture m, μ_m and Σ_m are the parameters of the Gaussian mixture component, m.

Expanding and ignoring the terms independent of warping factor α, estimation of a warping factor using this criteria can be shown to be equivalent to maximizing the following auxiliary function [3].

$$Q(\alpha) = \sum_{f=1}^{F} \sum_{m=1}^{M} \gamma_m \left[-\frac{1}{2}(A_\alpha x - \mu_m)^T {\Sigma_m}^{-1} (A_\alpha x - \mu_m) \right]$$
$$+ \beta \log|A_\alpha| + \log p(\alpha|\Theta) \text{ where, } \beta = \sum_{f=1}^{F} \sum_{m=1}^{M} \gamma_m$$

In the case of a single mixture β could reduce to F, the total number of frames. Optimizing this function requires the calculation of the matrix derivative. The form of the warping matrix renders inappropriate the CMLLR solution of decomposition of the determinant derivative using cofactors. A set of precomputed α matrices can be multiplied with the sufficient statistics to estimate the optimal warping factors [8]. This approach reduces to a grid search rather than gradient descent estimation. Higher order terms in the matrix can be ignored to give a closed form solution [14, 15]. Optimization using lower order terms in the matrix or using few lower order cepstral coefficients does not guarantee maximization of the auxiliary function for the entire feature length. This work presents Brent's search for finding the optimal value of the warping factor from this auxiliary function. Assuming a diagonal covariance for the auxiliary function results in minimization of the following function.

$$Q(\alpha) = \frac{1}{2} \sum_{f=1}^{F} \sum_{m=1}^{M} \gamma_m \sum_{i=1}^{N} \frac{(A_{\alpha_i} x_i - \mu_{m_i})^2}{\sigma^2_{m_i}} - \beta \log |A_\alpha| - \log p(\alpha|\Theta) \quad (7)$$

where, N is the dimensionality of the features. Brent's [16] search is used to find an optimal warping factor with this auxiliary function. For VTLN, the search is bounded using a bracket of -0.1 and 0.1.

The auxiliary function represented by EM can use the statistics as in CMLLR estimation [13]. It results in the following auxiliary function.

$$Q(\alpha) = \frac{1}{2} \sum_{i=1}^{N} (w_i G_i w_i^T - 2 w_i k_i^T) - \beta \log |A_\alpha| - \log p(\alpha|\Theta) \quad (8)$$

where,

$$G_i = \sum_{m=1}^{M} \frac{1}{\sigma^2_{m_i}} \sum_{f=1}^{F} \gamma_m x_f x_f^T \text{ And } k_i = \sum_{m=1}^{M} \frac{1}{\sigma^2_{m_i}} \mu_{m_i} \sum_{f=1}^{F} \gamma_m x_f^T$$

and w_i represents the i^{th} row of the transformation matrix A_α. Time complexity for 'E-step' of the VTLN EM optimization is same as CMLLR transform estimation ($O(n^3)$). But, the 'M-step' using Brent's search is only of the order $O(\log(n)M(n))$, compared to $O(n^4)$ for CMLLR.

4.3. Multiclass VTLN

VTLN is generally implemented using a single warping factor for an entire utterance or most often all the utterances of a single speaker representing a global spectral warping. All phonemes do not exhibit the same spectral variation due to physiological differences [17]. It should be more effective to use different warping factors for different phone classes. Multiple warping factors have yielded improvements in recognition performance. Data can be divided into acoustic classes using data-driven approach or using phonetic knowledge as shown in [17]. Phoneme dependent warping can be implemented after obtaining phone labels from a first pass recognition [18]. Frame specific warping factors can also be estimated by expanding the HMM state space with some constraints [19].

Different phone classes can be synthesized with different warping factors for a single speaker. Multiple transforms are usually applied using a regression class tree. Such regression classes can also be employed in multi-class VTLN. The regression class tree structure is derived from the decision tree clustering as in HTS [12]. Each regression class can have different warping factors. This can result in different warping for different classes resulting in appropriate warping for each sound as anticipated on factors like place of articulation. This research investigates the multi-class EM-VTLN estimation in the context of statistical synthesis. The issues with VTLN estimation for higher order features are investigated in the next section.

5. Challenges in Warping factor estimation

Problems with warping factor estimation using Jacobian normalization are discussed in [11]. These problems are further exacerbated by the higher order features. The following sections present warping factor estimation problems in these two scenarios. Ideal warping factors for synthesis using subjective evaluations are used to define the optimal technique for warping factor estimation from HMM.

5.1. Higher Order features

Higher order cepstral features will also capture aspects of spectral fine structure and are not limited to the spectral envelope. This causes problems when estimating the values of α. It can be observed from the Figure 1c that the warping factors concentrate on the middle of the range of the warping factors, which results in very little warping (or no warping) for many speakers. This effect is not seen in lower order features, which have a bimodal distribution for male and female speakers with a larger range of warping factors as shown in the Figure 1b.

EM formulation of bilinear transform warping is used to estimate the VTLN parameters shown in the figure. Experiments were also performed to confirm that this phenomena is not due to the local minima problem of the EM optimization. The warping factors for male and female speakers were initialized with the extreme values of 0.1 and -0.1 respectively. Even with this initialization, the warping factors converged to similar values as shown in the figure after few iterations of VTLN. In [2], the authors proposed initialization with warping factors estimated from lower order features as a work around for this problem.

It was also noted that there can be numerical instabilities while calculating the inverse of the transformation matrix for higher order features. This problem is addressed by generating the matrix using the inverse of the α value. $A_\alpha^{-1} = A_{\alpha^{-1}}$ This is an additional advantage of using bilinear transform.

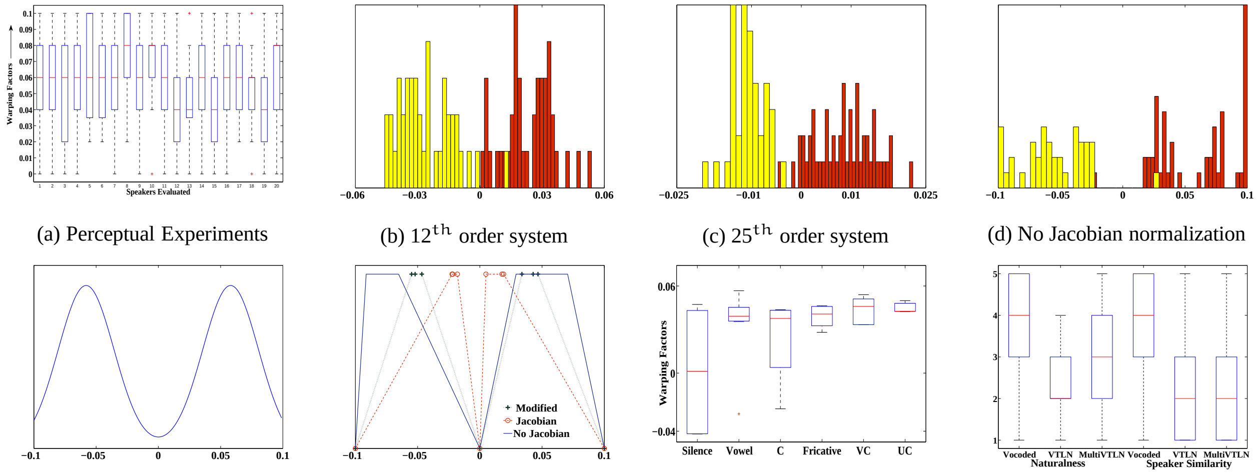
5.2. Jacobian Normalization

Estimation of warping factors using Jacobian normalization reduces the spread of α distribution and restricts the warping factor values to a small range of αs. Even though omitting the use of Jacobian normalization can estimate more distributed warping factors, it is observed that Jacobian normalization is important especially for higher order features. It can be seen from Figure 1d that the warping factors tend towards the boundaries when not using Jacobian normalization, thus resulting in an unstable estimation. Likelihood scaling or using prior probability are two techniques to improve the spread of warping factor values [11]. Increasing the spread of α distribution does not necessarily lead to better recognition performance in ASR [20].

5.3. Optimal warping factors for synthesis

The main motivation of the perceptual experiments presented in section 3 is to find the optimum distribution of α for TTS. One of the objectives of this paper is to find techniques that provide similar distributions from EM-based warping factor estimation. The resulting methods may not agree with the approaches previously proposed for ASR. It could be hypothesized that ML is not the right criterion for VTLN in TTS. Techniques like minimum generation error (MGE)[21] that have shown better results with HMM synthesis may perform better, but this paper focusses only on the ML criterion. The hypothesis from earlier results with grid search presented in [2] is that synthesis demands higher warping factors. Similar inference is made from preliminary experiments and the detailed subjective evaluations.

The challenge is that use of Jacobian normalization reduces the amount of warping. Ideally, not using Jacobian gives higher



(a) Perceptual Experiments

(b) 12th order system

(c) 25th order system

(d) No Jacobian normalization

(e) Prior from Perception

(f) Adaptation data

(g) Phoneme Classes

(h) Subjective Evaluation

Figure 1: Distributions over warping factor value. The abscissa in cases (b-f) is α, although note that the ranges vary.

values of α and gives a good spread to the distribution as expected in TTS. But, not using Jacobian is theoretically incorrect especially for higher order features. A few techniques were presented in [11] to increase the spread of α values. Scaling the log-likelihood score or using an appropriate prior distribution can show some improvements. The combination of these two techniques is expected to give better α values. Subjective evaluations present a good prior for the warping factor distribution for synthesis. The prior based on a mixture beta distribution estimated from the statistics obtained from the perceptual experiments is shown in Figure 1e.

Using a prior distribution derived from the perceptual experiments cannot have any effect without a scale factor. The non-use of Jacobian has a similar effect to the right prior with the right scale factor. These values can only be estimated empirically. Hence, it is pragmatically helpful to estimate the warping factors in the desired range without using Jacobian normalization. This is not true for regression class based VTLN estimation where each class has a separate warping factor. If Jacobian normalization is not used for multiple class VTLN, the warping factors go out of bounds for some classes. For these classes the EM auxiliary function is no longer convex and the Brent's search fails to find a minimum. The best available technique that could be used for multiple transform case is the combination of scaled likelihood and scaled prior with Jacobian normalization.

6. Evaluations with VTLN

The distribution of warping factors for the adaptation data is presented in Figure 1f. One utterance for each speaker from a subset of WSJCAM0 evaluation set is used as the adaptation data. It can be observed that the warping factor distribution using Jacobian normalization on the adaptation data does not give α values as expected for synthesis. The prior distribution obtained from the perceptual evaluations is closer to that obtained without using Jacobian normalization. The best method to increase the amount of warping is observed to be the combination of scaled likelihood and scaled prior. None of these schemes could achieve the spread that is obtained from not using Jacobian normalization. The warping factors for different speakers are very close and is not helpful in differentiating between them.

Even though single parameter cannot capture too many characteristics of the speaker, some preliminary results are shown with single and multi-class VTLN. The experimental setup presented in [2] is used for these experiments. VTLN is implemented as EM optimization embedded in the HMM training. VTLN is not comparable with other techniques like CMLLR since there are not enough parameters to represent most of the speaker characteristics. Ideally, as shown in [2], VTLN should give additive improvements to CMLLR. The current implementation does not combine the two techniques. Future research will be focussed on this task.

6.1. HMM Speech Synthesis

An average voice model built using WSJCAM0 is used to estimate the warping factors for the target speaker. Experiments are performed with single and multi-class transforms. A single utterance is used as adaptation data for all the techniques. Different techniques for estimating warping factors are evaluated using the objective measure based on mel-cepstral distortion (MCD). MCD is the Euclidean distance of synthesized cepstra with that of the values derived from the natural speech.

The results in Table 3 support the fact TTS demands a higher range of warping factors which can be achieved through not using Jacobian normalization or using a prior with likelihood scaling. The best method observed is the combination of Jacobian normalization with scaled prior and scaled likelihood (denoted as 'Modified'). It was observed that inter-speaker variability for α values is better when not using Jacobian normalization. Hence, the subjective evaluations are performed with α values estimated without using Jacobian for single class VTLN and using the combination of Jacobian with scaled prior and scaled likelihood for multi-class VTLN. It is observed during training that the log likelihood score of the training data improved consistently while using single and multi-class VTLN. This is also evident from the MCD scores shown in the table, which are higher for the SAT trained models than the SI average voice models. SI models are trained without any adaptation and SAT models are trained with multiple iterations of VTLN transformation and HMM parameter estimation. The multi-class SAT trained VTLN has maximum MCD and should represent the best average voice. Even with very little adaptation data, multi-class VTLN gives better MCD scores.

The distribution of α for different phoneme classes for a male speaker is shown in the Figure 1g. It is observed that silence has very noisy warping factors and ideally should be ignored in adaptation. Multi-class VTLN can facilitate this task by ignoring the classes representing silence. 'C' represents con-

Table 3: MCD (in dB) for VTLN synthesis. Label ”VTLN” represents the single parameter and ”Multi-VTLN” is the regression class based multiple transform VTLN. MCD for Average Voice (SI model) without using VTLN is 1.118

	VTLN	Multi-VTLN
Average Voice (SAT)	1.153	1.197
No Jacobian	1.014	-
Jacobian	1.080	1.062
Jacobian+Scaled_Prior	1.035	1.019
Jacobian+Scaled_LL	1.001	0.972
Modified	0.984	0.948

sonants in general with warping factors tending to lower values. The Voiced (‘VC’) and Unvoiced (‘UC’) category of consonants show somewhat opposite trends to each other. The values are derived from the Modified method of using Jacobian normalization with scaled likelihood and prior. The warping factors are slightly biased towards the prior for all classes which explains the high warping factors for some consonant classes. A clearer difference is observed in the case where no prior is used.

Speaker similarity and Naturalness are the subjective measures evaluated. Evaluations are performed on 60 sentences from 3 different systems. Systems evaluated are vocoded speech, single parameter global VTLN transforms and multiple transforms based VTLN using regression classes. The parameters are estimated using a single adaptation utterance. Jacobian normalization is not used in single transform case and Jacobian with scaled likelihood and prior is used in multi-VTLN case. Listeners were asked to rate the sentences on a 5 point scale, 5 being “completely natural” or “sounds exactly like speaker” and 1 being “completely unnatural” or “sounds like a totally different speaker”. 20 listeners participated in the evaluation and results are presented in Figure 1h. There is not much difference in speaker characteristics perceived using single or multiple transform VTLN, but the naturalness is a little better for the multiple transform case. This is a contradiction to the observations with CMLLR, which sounds less natural to VTLN [2]. The reason for this phenomenon could be that multiple transforms in effect is just a better implementation of VTLN and performs appropriate warping on different sounds. The difference could also be due to different techniques used in single and multi-class VTLN. Further investigation needs to be performed on this result. Even though the MCD values are not very far apart, (not as much as the values which are usually postulated as the perceivable change in speech), the difference is easily perceived in subjective evaluations. Demos available at `www.idiap.ch/paper/ssw7_vtln/demo.html`.

7. Conclusions

This work presents an efficient and accurate implementation of VTLN based on EM. Appropriate warping factors for TTS are analyzed and techniques are suggested to estimate similar values from the model. Regression class based multiple transform VTLN is also presented which performs appropriate warping on different sounds. VTLN has a limited number of parameters (single warping factor in case of bilinear transform) to be estimated. On one hand, this enables estimation of warping factors and adaptation using very little adaptation data. On the other hand, there is only limited characteristics that this parameter can capture. VTLN is not comparable to model based transformations like CMLLR especially when there is large amount of adaptation data. Hence, in order to get improvements in adaptation when more adaptation data is available, VTLN should be combined with CMLLR. In order to combine VLTN with CMLLR, the current research focusses on implementing VTLN as a prior to CMLLR transform as in constrained structural maximum a posteriori linear regression (CSMAPLR).

8. Acknowledgements

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

9. References

[1] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, “Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 66–83, 2009.

[2] 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.

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

[4] 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.

[5] L. Lee and R. Rose, “A frequency warping approach to speaker normalization,” *IEEE Transactions on Speech and Audio Processing,*, vol. 6, pp. 49–60, 1998.

[6] M. Pitz and H. Ney, “Vocal tract normalization equals linear transformation in cepstral space,” *IEEE Transactions on Speech and Audio Processing,*, vol. 13, pp. 930–944, 2005.

[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] 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.

[9] 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 *Proceedings of Interspeech*, Brisbane, Australia, 2008, pp. 1713–1716.

[10] A. Sankar and C.-H. Lee, “A maximum-likelihood approach to stochastic matching for robust speech recognition,” vol. 4, no. 3, pp. 190–202, May 1996.

[11] L. Saheer, P. N. Garner, and J. Dines, “Study of Jacobian normalization for VTLN,” *Idiap-RR-25-2010*, 2010. [Online]. Available: http://publications.idiap.ch

[12] J. Yamagashi, H. Zen, T. Toda, and K. Tokuda, “Speaker independent HMM based speech synthesis system - HTS-2007 system for blizzard challenge 2007,” in *Proceedings of Sixth ISCA Workshop on Speech Synthesis*, Bonn, Germany, 2007.

[13] M. J. F. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” *Computer Speech & Language,*, vol. 12 (2), pp. 75–98, 1998.

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

[15] 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.

[16] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in C*. Cambridge University Press, 1992.

[17] S. P. Rath and S. Umesh, “Acoustic class specific VTLN-warping using regression class trees,” in *Proceedings of Interspeech*, Brighton, UK, 2009, pp. 556–559.

[18] S. Molau, S. Kanthak, and H. Ney, “Efficient vocal tract normalization in ASR,” in *Proceedings of ESSV*, Cottbus, Germany, 2000.

[19] A. Miguel, E. Lleida, R.L.Buera, and A. Ortega, “Augemented state space acoustic decoding for modeling local variability in speech,” in *Proceedings of Interspeech*, Lisbon, Portugal, 2005.

[20] M. Pitz, “Investigations on linear transformations for speaker adaptation and normalization,” Ph.D. dissertation, RWTH Aachen University, 2005.

[21] Y.-J. Wu and R.-H. Wang, “Minimum generation error training for HMM-based speech synthesis,” in *Proceedings of ICASSP*, France, 2006, pp. 89–92.