Vocal Tract Length Normalization for Statistical Parametric Speech Synthesis

Lakshmi Saheer, *Student Member, IEEE,* John Dines, *Member, IEEE,*
and Philip N. Garner, *Senior Member, IEEE*

Abstract

Vocal tract length normalization (VTLN) has been successfully used in automatic speech recognition for improved performance. The same technique can be implemented in statistical parametric speech synthesis for rapid speaker adaptation during synthesis. This paper presents an efficient implementation of VTLN using expectation maximization and addresses the key challenges faced in implementing VTLN for synthesis. Jacobian normalization, high dimensionality features and truncation of the transformation matrix are a few challenges presented with the appropriate solutions. Detailed evaluations are performed to estimate the most suitable technique for using VTLN in speech synthesis. Evaluating VTLN in the framework of speech synthesis is also not an easy task since the technique does not work equally well for all speakers. Speakers have been selected based on different objective and subjective criteria to demonstrate the difference between systems. The best method for implementing VTLN is confirmed to be use of the lower order features for estimating warping factors.

Index Terms

Vocal tract length normalization, Expectation maximization optimization, HMM based statistical parametric speech synthesis, Speaker adaptation

I. Introduction

The ability to transform voice identity in text-to-speech synthesis (TTS) is an important area of research with applications in medical, security and entertainment industries. One specific application that has seen considerable interest by the research community is that of personalized speech-to-speech translation. It is crucial to this kind of application that the speaker characteristics are induced into the output speech from the very first utterance spoken by a speaker. Thus, speaker characteristics need to be estimated from very little adaptation data.

The authors are with Idiap Research Institute, Switzerland. L. Saheer is also affiliated with Ecole Polytechnique Fédérale de Lausanne, Switzerland. e-mail: (lsaheer,dines,pgarner)@idiap.ch.

Statistical parametric synthesis [1] using hidden Markov models (HMM) has proven to be a particularly flexible and robust framework for performing speaker transformation, leveraging off a range of speaker adaptation techniques previously developed for automatic speech recognition (ASR) [2]. The maximum likelihood linear transformation (MLLT) based adaptation techniques entail linear transformation of the means and variances of an HMM to match the characteristics of the speech for a given speaker. Feature adaptation, on the other hand, transforms the feature vectors rather than the model parameters. It is evident that linear feature transformations, such as constrained maximum likelihood linear regression (CMLLR [3]), also have a straight-forward model adaptation formulation. Aside from MLLT-based feature transformation, there are a number of feature adaptation methods that employ considerably simpler adaptation schemes based on only a few parameters. Such feature adaptation schemes can be carried out with very little adaptation data.

Vocal tract length normalization (VTLN) [4] is one such rapid speaker adaptation (or *normalization*) technique that is widely used in ASR. 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. The same technique can likewise be used to transform the average voice to match the VTL of a target speaker.

Initial investigations of VTLN for statistical parametric speech synthesis were reported by Saheer *et.al.* [5] using a grid search and later on, presented its efficient implementation using expectation-maximization (EM) with Brent's search [6]. In this paper we further elaborate the EM implementation of VTLN, provide a detailed analysis of perception and evaluation of VTLN, followed by comprehensive subjective evaluations.

While drawing on considerable work that has previously been conducted in the application of VTLN for ASR, we note that this paper presents a number of novel contributions in applying VTLN for TTS. These contributions centre around the significant differences between feature extraction methods typically applied in ASR and TTS. Statistical speech synthesis uses mel-generalized cepstral coefficients (MGCEP), while ASR usually uses mel frequency cepstral coefficients (MFCC) as features. In contrast to MFCC, the absence of filter-bank analysis in combination with bilinear mel-warping in MGCEP enable the novel implementation of a zero overhead VTL warping (by warping the untruncated cepstrum with the negative warp factor). Work presented in this paper is build upon this novel idea presented in an earlier paper [5].

Statistical speech synthesis also tends to employ high dimensionality feature analysis to achieve desired synthesis quality. Such high dimension features can pose problems for effective VTLN, as was shown

by Saheer *et.al.* [6]. In particular, application of Jacobian normalization becomes a critical factor [7] that has often not been conclusively dealt with in ASR studies (some suggest it should be ignored, others the contrary). This paper proposes new techniques to overcome these problems and further investigates earlier approaches to find the technique that achieves the best performance for synthesis.

Earlier EM based VTLN implementations either used a grid of values or ignored the complicated terms in the auxiliary function to estimate a gradient. This work presents Brent's search based EM solution to VTLN warping factor estimation. Since VTLN is a highly constrained transformation, it is not expected to reproduce the full range of voice qualities associated with more powerful MLLT-based adaptation techniques. This makes evaluation of VTLN non-trivial since its effectiveness can greatly vary from speaker to speaker. This paper presents a detailed study on how to select an appropriate speaker for subjectively evaluating VTLN and presents more comprehensive results than previously reported.

In this paper, we do not intend to present VTLN as an alternative adaptation technique in comparison with other linear transformations. This work addresses the issues associated with implementation of VTLN for statistical speech synthesis. As in ASR (also shown in the earlier work [5]), VTLN is expected to give additional performance improvements to other linear transformations. VTLN as such is an efficient technique in voice conversion applications, (especially when there is no explicit target speaker) and hence it is very important to find the right method to apply it to the speech synthesis domain.

The paper is organised as follows. The next section gives a background on VTLN as used in ASR and presents the VTLN formulation that can be used in synthesis. The details on the EM formulation for VTLN is presented in Section III followed by the challenges faced in synthesis using VTLN in section IV along with proposed solutions. This section also presents the optimal warping factors useful for perception and the challenges in selecting optimal speakers for evaluating VTLN. Details of the subjective and objective evaluations and results are presented in section V with conclusions in section VI.

II. Background

Vocal tract length normalization is a widely used technique in ASR, made particularly attractive due to its simplicity and robustness. One of the main advantages of VTLN is that it is typically characterized by only one parameter (the warping factor), which can be reliably estimated even with a single adaptation sentence. While such advantages may equally be of interest in TTS, there have been very few attempts to date towards this end. This section presents some details of the earlier work in this field, and hence motivates the research presented in this paper.

A. VTLN for Automatic Speech Recognition

VTLN normalizes the position of the formant peaks by warping the spectrum to represent an average vocal tract. The components involved in this technique are:

- A family of warping functions (linear, piecewise linear, non-linear, bilinear, etc.)
- A warping factor (α for bilinear transform)
- An optimization criterion (MAP, ML, MGE, etc.)

The main advantage of VTLN is that very little adaptation data is required to estimate reliable warping factors. Optimal parameters for the warping function referred to as warping factors are estimated based on a pre-determined optimization criterion. The warping factors are usually selected from a grid of available values. This technique is referred to as "grid search" in this paper and involves high time and space complexity due to the extraction of features for each speaker using each warping factor in the grid.

Spectral transformations are closely tied to the underlying feature analysis technique, hence, implementation of VTLN cannot be easily separated from the feature extraction stage. An initial implementation of VTLN for ASR by Lee and Rose [4] was carried out in the framework of MFCC features. An efficient VTL warping was implemented by modifying the filterbank computation. The optimal warping factor for a piecewise linear warping function was estimated using a maximum likelihood (ML) based optimization criterion given by:

$$\hat{\alpha}_s = \underset{\alpha_s}{\operatorname{argmax}}\, p(\mathbf{x}_{\alpha_s} \mid \mathbf{\Theta}, w_s) \tag{1}$$

where $\mathbf{x}_{\alpha_s}$ represents the features warped with the warping factor α_s, which is the warping factor for speaker s. $\mathbf{\Theta}$ 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. This expression for feature transformation ignores the Jacobian normalization term usually implemented as the determinant of the transformation matrix. The simultaneous updation of features and calculation of likelihood scores is not a very consistent method for performing VTLN.

Though it is an inexact method for performing VTLN, if the warped spectrum could be estimated from the unwarped cepstral features, the full feature computation for each point in a grid search could be bypassed. The development of cepstrum domain VTLN [8, 9, 10] helped alleviate this issue further by demonstrating the equivalence of spectral warping and linear cepstrum transformations [11]. VTLN has also been implemented in other feature domains such as perceptual linear prediction (PLP) features [12] for ASR. A break-through in this field was the expectation-maximization (EM) formulation [10], which improved efficiency through the use of a gradient descent search (on a grid of warping factor values). Most of these implementations still used a grid of pre-computed transformation matrices or a simple (linear) warping function to estimate a tractable gradient for the EM auxiliary function.

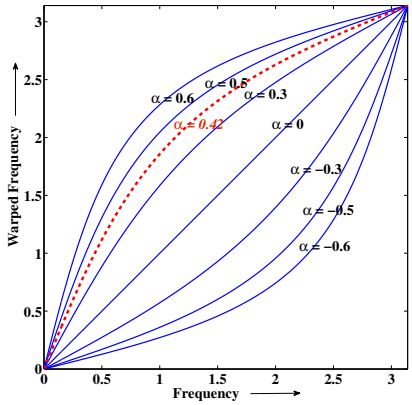


Fig. 1: The bilinear Transform. Setting $\alpha = 0.42$ approximates mel-frequency scale at 16kHz sampling rate.

The all-pass transform can be used to approximate most commonly used transformations in VTLN [12, 13]. The all-pass transform based VTLN introduced a broader range of possibilities for transformation of the spectra in the ASR domain. The bilinear transform based warping function, which is the simplest form of the all-pass transforms (shown in Figure 1), has only a single variable α as the warping factor. This parameter is representative of the ratio of the vocal tract length of the speaker to an average vocal tract length. The terms warping factor and 'α' refer to the same parameter and are used interchangeably throughout this paper.

B. VTLN in Text to Speech Synthesis

Though widely used in ASR, VTLN is a comparatively less exploited technology in TTS. Earlier attempts at using VTLN for concatenative speech synthesis [14] had the advantage of performing voice conversion without needing to use corresponding time frames from the source and target speakers. This method was particularly useful in cross-language voice conversions where natural time alignments across speech from different languages could not be obtained. Voice conversion is performed by clustering the phoneme classes of source and target speakers separately, and then statistically mapping the phoneme classes across speakers. This method of voice conversion is non trivial and involves various parameters to be tuned such as the distance measure and minimization criterion, and also requires parameter smoothing techniques. With the advent of HMM-based statistical speech synthesis, voice conversion techniques, including VTLN, can be implemented in a more subtle manner [2].

1) Mel-generalized Cepstral Coefficients: MGCEP [15] features are one of the most widely used features for statistical speech synthesis. The steps involved in the MGCEP feature analysis are shown in Figure 2. Note that MGCEP is different from the traditional MFCC features most commonly used in ASR, in particular, there is no filterbank analysis. A generalized log function is applied on the squared

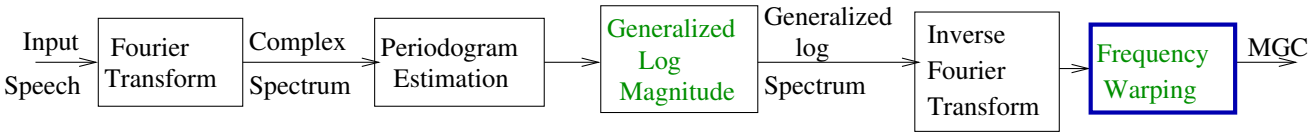


Fig. 2: Key stages of mel-generalized cepstral analysis. UELS and cepstral truncation are omitted. Spectral warping is implemented as a linear transformation in the cepstral domain (as shown).

magnitude spectrum and the mel-warping is performed in the cepstral domain using the all-pass filter based bilinear transform.

The generalized cepstral analysis method can be viewed as a unified approach to the cepstral and linear prediction methods, in which the model spectrum varies continuously from all-pole to cepstral. The generalized logarithmic function is a natural generalization of the logarithmic function with a parameter γ:

$$s_\gamma(\omega) = \begin{cases} \frac{\omega^\gamma - 1}{\gamma}, & 0 < |\gamma| \leqslant 1, \\ \log \omega, & \gamma = 0. \end{cases} \tag{2}$$

where, s is the generalized function and ω represents the magnitude spectrum. The warping function used in this feature extraction technique is the all-pass transform based bilinear warping with warping parameter α. This frequency transformed generalized cepstrum has frequency resolution similar to that of the human ear with an appropriate choice of the value of the warping parameter(α). Hence, it is expected that the mel-generalized cepstral coefficients are useful for speech spectrum representation.

2) VTLN for HMM-based Synthesis: There have been earlier attempts to use VTLN for building better speaker adaptive models [16] for statistical speech synthesis. The first attempt to use VTLN in statistical speech synthesis to transform target voices was by Saheer et al [5]. In this work, MGCEP features were used with the bilinear transform based VTLN. Warping parameter estimation was carried out with ML optimization over a grid search. It was shown that VTLN brings in some speaker characteristics and provides additional improvements to constrained maximum likelihood linear regression (CMLLR), especially when there is a limited number of adaptation utterances.

C. Motivation for the current work

Any general VTLN implementation faces two main challenges, (1) efficient calculation of VTL-warped features for a given warping factor and (2) optimization of the warping factor with low time and space overhead. In addition, application of VTLN for TTS requires addressing of issues specific to this domain.

The first of these issues can be addressed by implementing VTLN as a cepstral transformation. It is argued persuasively by Pitz and Ney [11] 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 [15], with

the use of bilinear warping function being of particular interest due to its presence in the mel-frequency warping function of MGCEP features.

The second issue is addressed through formulation of warping factor estimation in the framework of expectation maximization. Representation of VTLN as a model transformation enables the use of EM for finding the optimal warping factors [10, 17]. This provides advantages in terms of more precise estimation of warping factor, α, and improved efficiency in time and space. 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.

Initial implementation of these techniques along with additional challenges faced in statistical synthesis have been presented in earlier work [6]. This paper gives a more detailed description of the ideas presented in previous work with additional evaluations and results. This paper also presents in detail the challenges faced in implementing and evaluating VTLN for statistical speech synthesis along with proposed solutions.

III. Theory

A. VTLN in Statistical Parametric Speech Synthesis

Speaker adaptive statistical speech synthesis most commonly uses MGCEP features, which incorporate a bilinear transform to achieve mel-scale frequency warping. In this work, VTLN is also implemented as a bilinear transform warping with ML optimization of the warping factor, α. The warping performed [18] by the bilinear transform of a simple first order all-pass filter with unit gain is shown in Figure 1.

Bilinear transform based VTLN can be combined with MGCEP as the concatenation of two bilinear transforms which are equivalent to a single bilinear warping with warping factor:

$$\alpha = \frac{\alpha_V + \alpha_M}{1 + \alpha_V \cdot \alpha_M} \tag{3}$$

where, α_M and α_V represent the warping factors for MGCEP features and VTLN respectively. It has been observed that a value of $\alpha = 0.42$ can approximate mel-scale at a sampling rate of 16kHz. Bilinear VTL warping thus, in principle, constitutes zero additional overhead (by warping the untruncated cepstra with the negative warp factor) with respect to the standard MGCEP feature analysis during synthesis.

B. Expectation Maximization Formulation

It has been shown that EM can be used to estimate VTLN warping factors for ASR [12, 10, 17]. 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 [3]. 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 by Saheer *et.al.* [5]. Grid search has two notable drawbacks: The first is that the granularity of the grid restricts the precision of the warping factor. The second is that the likelihood estimation for many candidate warping factors is computationally expensive.

This work presents an EM formulation for the warping factor estimation which improves upon the 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.

1) VTLN as model transform:

The bilinear transform of a simple first-order all-pass filter with unit gain leads to a warping of the frequency ω into $\tilde{\omega}$ in the complex z-domain as follows:

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \qquad |\alpha| < 1 \tag{4}$$

where $z^{-1} = e^{-j\omega}$, $\tilde{z}^{-1} = e^{-j\tilde{\omega}}$, and α is the warping factor. As shown by earlier work [11, 12], linear transformation in the cepstral domain c_k for the bilinear transform yields a warped cepstrum ($\tilde{c}_m$) given by:

$$\tilde{c}_m = \sum_k A_{mk}(\alpha)\, c_k \tag{5}$$

where $A_{mk}(\alpha)$ is the m-th row k-th column element of the warping matrix $\mathbf{A}_\alpha$ consisting of the warping factor α and the Cauchy integral formula yields [11]:

$$A_{mk}(\alpha) = \frac{1}{(k-1)!} \sum_{n=\max(0,k-m)}^{k} \binom{k}{n} \frac{(m+n-1)!}{(m+n-k)!} (-1)^n \alpha^{2n+m-k}. \tag{6}$$

We may represent the linear transformation in the vector form

$$\mathbf{x}_\alpha = \mathbf{A}_\alpha \mathbf{x} \tag{7}$$

where $\mathbf{x}_\alpha = (\tilde{c}_1, \cdots, \tilde{c}_M)^\top$ and $\mathbf{x} = (c_1, \cdots, c_K)^\top$. The transform may also be directly applied to the dynamic features of the cepstra; the transformation matrix is block diagonal with repeating $\mathbf{A}_\alpha$ matrix.

$$\mathbf{x}_\alpha = \begin{bmatrix} \mathbf{A}_\alpha & 0 & 0 \\ 0 & \mathbf{A}_\alpha & 0 \\ 0 & 0 & \mathbf{A}_\alpha \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \Delta\mathbf{x} \\ \Delta^2\mathbf{x} \end{bmatrix} \tag{8}$$

where, $\mathbf{x}_\alpha$ is the warped cepstral coefficients, $\mathbf{x}$ is the static features, $\Delta\mathbf{x}$ and $\Delta^2\mathbf{x}$ are dynamic part of the

cepstra. The unwarped cepstral features are multiplied with the linear transformation matrix to generate warped features. This results in significant computation savings since features need not be individually recomputed for each warping factor. Equation 6 can be represented in the form of a recursion [15], which in turn, can be calculated as a transformation matrix. This matrix representation of the MGCEP bilinear transform in the cepstral domain was presented (with a minor error) in [5]:

$$\mathbf{A}_\alpha = \begin{bmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{M_1} \\ 0 & 1-\alpha^2 & 2\alpha(1-\alpha^2) & \cdots & M_1\alpha^{M_2}(1-\alpha^2) \\ 0 & -\alpha(1-\alpha^2) & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & (-1)^M(1-\alpha^2)\alpha^{M_2} & \cdots & \cdots & \cdots \end{bmatrix}$$

where $M_1 = M - 1$ and $M_2 = M - 2$.

In practice, there may be small differences in the resulting features depending upon where the warping is performed. For instance, the UELS (unbiased estimate of the log spectrum) iteration used by MGCEP will introduce a small distortion from spectrum to cepstrum. Also, transformation in the feature extraction stage will operate on a full cepstrum, but at the model stage will act on a truncated cepstrum. The result is that, whilst the resulting transform (Equation 5) is valid, it may actually represent the sum of these distortions rather than simply VTL.

Similar to CMLLR adaptation, the feature transform can be analogously represented as a model transform [3]. The maximum likelihood optimization for a Gaussian distribution in the feature domain is:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}} |\boldsymbol{A}_\alpha| \mathrm{N}(\boldsymbol{A}_\alpha \boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{9}$$

The same equation can be represented as a model transform:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}} \mathrm{N}(\boldsymbol{x} \mid \boldsymbol{A}_\alpha^{-1}\boldsymbol{\mu}, \boldsymbol{A}_\alpha^{-1}\boldsymbol{\Sigma}(\boldsymbol{A}_\alpha^{-1})^{\mathsf{T}}) \tag{10}$$

where μ and Σ correspond to the mean and variance of the Gaussian model. The Jacobian normalization can be calculated as the determinant of the matrix $\boldsymbol{A}_\alpha$ representing the linear transformation of the cepstral features.

2) EM Auxiliary function: The maximum a posteriori (MAP) criteria for warping factor estimation similar to the ML criteria introduced in Equation 9 can be represented as follows:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}} |\boldsymbol{A}_\alpha| \mathrm{p}(\boldsymbol{x}_\alpha \mid \boldsymbol{\Theta}, w) \mathrm{p}(\alpha \mid \boldsymbol{\Theta}) \tag{11}$$

where, $\mathrm{p}(\alpha \mid \Theta)$ is the prior probability of α for a given model, $\boldsymbol{\Theta}$. When using the likelihood comparison to search for the best warping factor, Jacobian normalization has to be taken into consideration [8, 19]. The EM formulation of warping factor estimation results in the following auxiliary function. Taking the

log of the function and considering the Gaussian assumption for the model and marginalizing out the state sequences (hidden variable).

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}} \left\{ \sum_{f=1}^{F} \sum_{m=1}^{M} \gamma_{m,f} \Big[\log(N(\mathbf{A}_\alpha x_f \mid \mu_m, \Sigma_m)) + \log|\mathbf{A}_\alpha| \Big] + \log p(\alpha \mid \Theta) \right\} \quad (12)$$

where, $\mathbf{A}_\alpha$ is the transformation matrix for input feature vector $\mathbf{x}$, M is the total number of mixtures, F is the total number of frames, γ_m is the posterior probability of mixture m, and μ_m and Σ_m are the parameters of the Gaussian mixture component, m.

Expanding the terms in the above equation, estimation of a warping factor using this criteria can be shown to be equivalent to maximizing the following auxiliary function [12].

$$Q(\alpha) = \sum_{f=1}^{F} \sum_{m=1}^{M} \gamma_{m,f} \log \Big(\frac{1}{\sqrt{(2\pi)^N |\Sigma_m|}} \exp -\frac{1}{2} (\mathbf{A}_\alpha \mathbf{x}_f - \mu_m)^T \Sigma_m{}^{-1} (\mathbf{A}_\alpha \mathbf{x}_f - \mu_m) \Big) + \beta \log|\mathbf{A}_\alpha| + \log p(\alpha \mid \Theta) \quad (13)$$

where, N is the dimensionality of the features and,

$$\beta = \sum_{f=1}^{F} \sum_{m=1}^{M} \gamma_{m,f}$$

In the case of a single mixture, β could reduce to F, the total number of frames. Further expanding and setting the terms independent of warping factor α to K,

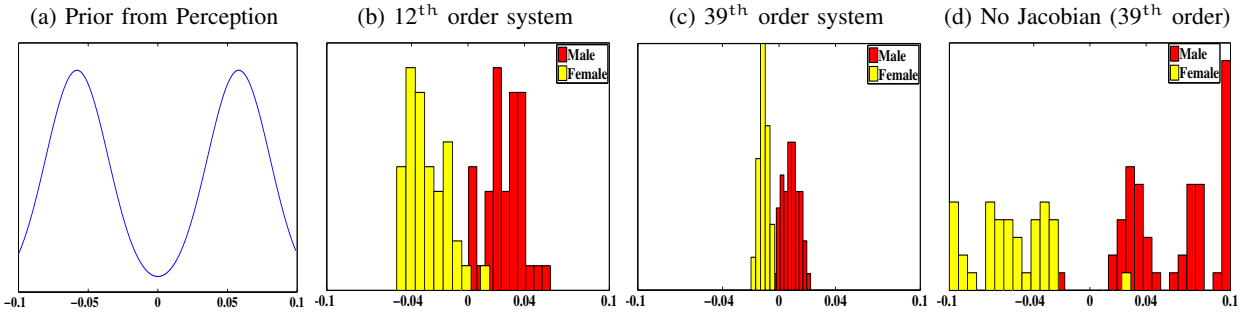
$$Q(\alpha) = \sum_{f=1}^{F} \sum_{m=1}^{M} \gamma_{m,f} \left[-\frac{1}{2} (\mathbf{A}_\alpha \mathbf{x}_f - \mu_m)^T \Sigma_m{}^{-1} (\mathbf{A}_\alpha \mathbf{x}_f - \mu_m) \right] + \beta \log|\mathbf{A}_\alpha| + \log p(\alpha \mid \Theta) + K \quad (14)$$

Assuming a diagonal covariance for the auxiliary function results in maximization of the following function.

$$Q(\alpha) = -\frac{1}{2} \sum_{f=1}^{F} \sum_{m=1}^{M} \gamma_{m,f} \sum_{i=1}^{N} \frac{(\mathbf{A}_{\alpha_i} \mathbf{x}_{f_i} - \mu_{m_i})^2}{\sigma_{m_i}^2} + \beta \log|\mathbf{A}_\alpha| + \log p(\alpha \mid \Theta) + K \quad (15)$$

The form of the matrix does not allow a closed form solution. A set of precomputed α matrices can be multiplied with the sufficient statistics to estimate the optimal warping factors [10]. 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 [20, 16]. Optimization using simple terms in the matrix or using few lower order cepstral coefficients does not guarantee that the estimated α will maximize the likelihood over the entire feature vector. We note that the dimensionality of order four (used by Hirohata

Fig. 3: Distributions over warping factor value. The abscissa is α, the warping factor. Ordinate represents the frequency of α.



et.al. [16]) is not sufficient to represent the spectral envelope which needs to be warped using VTLN (see also Figure 6) .

In this work, Brent's search [21] is used to find an optimal warping factor for the EM auxiliary function of Equation 15. Brent's search is a method for general one-dimensional root finding. This method combines root bracketing, bisection and inverse quadratic interpolation to converge from the neighbourhood of zero crossing. The main advantage of using Brent's search is that no derivative needs to be computed. For VTLN, the bracket of the search is bounded by -0.1 and 0.1.

The auxiliary function represented by EM (Equation 15) can use statistics similar to the CMLLR estimation as derived by Gales [3]. It results in the following auxiliary function.

$$Q(\alpha) = -\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}_i \mathbf{G}_i \boldsymbol{w}_i^T - 2\boldsymbol{w}_i \mathbf{k}_i^T) + \beta \log|\mathbf{A}_\alpha| + \log p(\alpha \mid \boldsymbol{\Theta}) + K \tag{16}$$

where,

$$\mathbf{G}_i = \sum_{m=1}^{M} \frac{1}{\sigma_{m_i}^2} \sum_{f=1}^{F} \gamma_{m,f} \boldsymbol{x}_f \boldsymbol{x}_f^T \tag{17}$$

$$\mathbf{k}_i = \sum_{m=1}^{M} \frac{1}{\sigma_{m_i}^2} \boldsymbol{\mu}_{m_i} \sum_{f=1}^{F} \gamma_{m,f} \boldsymbol{x}_f^T \tag{18}$$

and $\boldsymbol{w}_i$ represents the i^{th} row of the transformation matrix $\mathbf{A}_\alpha$.

3) Full MAP Estimation: The EM formulation presented here is a novel technique based on the MAP optimization criterion shown in Equation 11 instead of the ML optimization shown in Equation 1. There is a prior term $p(\alpha \mid \Theta)$ which determines the probable distribution of the warping factors depending on the model set. A general approach is to ignore this term or set it to unity resulting in an assumption of uniform distribution. While training, this term may not have any effect since there is a lot of data available for training and the data itself determines the underlying distribution of the warping factors. During testing, when the warping factor is to be estimated from a single utterance, this term could be

more valuable. In the case of synthesis, the challenge is to estimate the desired distribution of the warping factors which can be perceived as acceptable.

4) 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. Multiple warping factors have yielded improvements in recognition performance. Data can be divided into acoustic classes using data-driven approaches or using phonetic knowledge as shown by Rath and Umesh [22]. Phone dependent warping can be implemented after obtaining phone labels from a first pass recognition [23]. Frame specific warping factors can also be estimated by expanding the HMM state space with some constraints [24].

In speech synthesis, phone classes can also be synthesized with different warping factors for a single speaker. Multiple MLLTs 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 [25]. Each regression class can have different warping factors. This can result in different warping for different classes resulting in appropriate warping for each sound according to factors like place of articulation. This research investigates multi-class EM-VTLN estimation in the context of statistical synthesis. Unlike ASR, there are issues with VTLN estimation for statistical speech synthesis which are investigated in the next section.

C. Evaluation of VTLN as a speaker adaptation technique

As noted earlier, there is limited literature concerning application of VTLN in speech synthesis. The evaluation of VTLN in this context is in itself a research question. While this work is not set out explicitly to answer such a question, nonetheless some preliminary investigations are performed in this direction. In particular, this paper presents original work concerning the following points:

- How VTL transformations are perceived by humans, especially with respect to speaker similarity (Section IV-B1).
- Are there any acoustic correlates of perceived VTL (Section IV-B1)?
- What distribution of VTL warping factors are perceived as covering the range of human variation (Section IV-B1)?
- How VTLN can be effectively utilized in statistical speech synthesis framework (This section and Section IV-A).

IV. Challenges for VTLN synthesis

In this section, the main issues for the successful application of VTLN for statistical parametric speech synthesis are analysed. In particular, the challenges that are distinct from those encountered in past work

concerning VTLN, (most notably distinct from its application to ASR) are analysed in detail. These challenges centre around two main factors, numerical / computational and subjective / perceptual.

First of all, the impact of numerical modelling factors, particularly the impact of feature extraction for synthesis on VTL estimation is analysed. It is evident that ASR and TTS use feature analysis methods that differ quite markedly and this can have an impact on VTLN. In particular, the high feature dimensionality can pose significant difficulties. While the truncation of the cepstrum and the truncation of the transformation matrix can add additional challenges in the calculation of inverse transformation during synthesis. Thus, the challenges that TTS features introduce to VTLN are examined.

Second factor is that evaluation of TTS is conducted through subjective testing, hence, understanding of human perception and design of appropriate tests are necessary. In this section, some preliminary analysis on perception of vocal tract length vis-à-vis speaker similarity and whether or not there exist any correlates with low-level acoustic features are presented. Some thoughts on appropriate means to evaluate VTLN as a speaker adaptation / transformation approach in statistical parametric speech synthesis are also presented.

A. Modelling factors

The main differences between VTLN for ASR and TTS are in the feature type and feature dimensionality. Usually, MFCC features of the order of 12 are used in speech recognition. This provides a coarse representation of the spectral envelope and in turn the formant structure in the speech. The MGCEP features used for statistical speech synthesis have very high dimensionality — of the order of 25 or 39 — when compared to ASR features in order to obtain good synthesis quality by maintaining the fine structure of the spectrum.

VTL is normally considered as being related to formant location which in turn is captured sufficiently by lower order cepstra. Despite this, the higher order cepstra can still have a significant (and potentially detrimental) impact on warping factor estimation since they will make a significant contribution to the likelihood score calculation.

As an example of this phenomenon, it can be observed from the Figure 3c that the range of warping factors estimated from high order feature analysis is extremely narrow compared to those (shown in Figure 3b) for a more typical feature analysis order. In reality, such a narrow range of warping factors will be imperceptible in most cases. By contrast, it can be seen from Figure 3d that the omission of Jacobian normalization (to be discussed further below) will also result in warping factors far from what is expected to be appropriate, with divergence towards the boundaries.

In the remainder of this section various means are discussed to pragmatically address these issues as they relate to warping factor estimation for TTS.

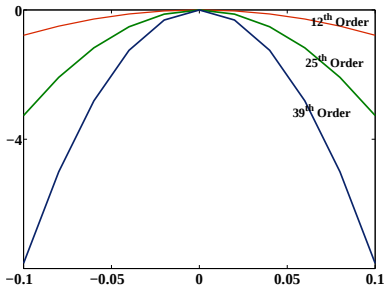


Fig. 4: Jacobian calculated as $\log|\mathbf{A}|$ for various feature dimensions. The abscissa is α, the warping factor and the ordinate is value of the Jacobian determinant.

1) Jacobian Normalization: Jacobian normalization should be applied to likelihood score calculations following a feature transformation (see appendix for derivation). In spite of this, it has been observed in ASR studies that use of Jacobian normalization tends to have an adverse effect 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 [13, 4, 26]. One recent study [27] on mismatched train and test conditions addressed this issue by compensating with a variance adaptation on top of VTLN. This approach is effectively able to compensate for the change in feature variability caused by the VTL transform, which is otherwise compensated for by Jacobian normalization.

The effect of Jacobian normalization is to penalize severe warping factors, thereby reducing the range of α that would otherwise be estimated in its absence. This is necessary since the warping transformation itself results in narrower variance of the transformed features that would result in an unfair bias towards extreme warping factors. This effect can be visualized in Figure 4 for different feature orders. It is evident that the Jacobian normalization term becomes more significant as the feature order increases and cannot be ignored as has been done in previous ASR studies.

It is apparent that the inclusion or omission of Jacobian normalization does not give desirable warping factors, hence, additional measures are required, as presented below.

2) Use of a prior distribution of warping factors: The prior on the warping factor in equation 25 has been omitted from VTLN implementations presented in the literature. Ignoring a prior normally corresponds to assuming a flat prior. Where sufficient data is available, this is often a reasonable approach. Conversely, when little data is available use of a prior can be important. We can assume that the prior on the warping factor should not be flat, more precisely:

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

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 3b,

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

$$p_{\propto}(\alpha \mid \mathbf{D}) \sum_{g \in \{m,f\}} (1+\alpha)^{p_g - 1}(1-\alpha)^{q_g - 1}, \tag{19}$$

where $\{p_m, q_m\}$ and $\{p_f, q_f\}$ are the pairs of beta parameters for male and female speech respectively, as in Figure 3a.

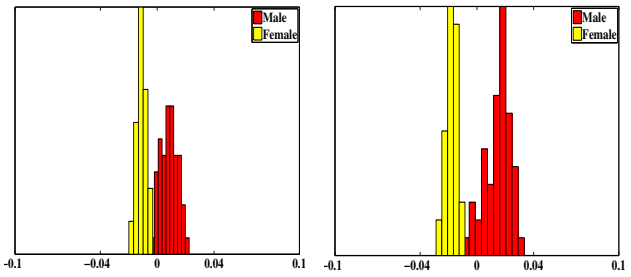
Note that omitting the Jacobian determinant from the likelihood calculation has the effect of using a prior with a PDF proportional to the inverse of the Jacobian (c.f. Figure 4). It biases α away from zero, enhancing the relative separation of the male and female modes. It can be seen from Figure 5a that the prior does not have much impact on warping factor estimation for the training data due to availability of sufficient 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 omitting Jacobian normalization improves performance especially in testing: during testing there is often insufficient data to generate a reliable estimate of warping factor, thus means to increase the spread of warping factors by omitting the Jacobian term acts as a reasonable prior in the case of low dimensional feature analysis.

3) Using likelihood scaling: In large vocabulary ASR, it is common to use a language model scale factor that in fact compensates for under estimation of acoustic likelihoods. This in turn is necessary because successive acoustic frames have much higher correlation than is captured by the HMM. Applied more correctly to the acoustic calculation, we might expect that the likelihood correction should apply to the likelihoods, but not to the Jacobian. In fact, this was investigated by Pitz [8], who applied the factor to the Jacobian analogous to the language model scale. This suggests an estimator of the form similar to the Equation 13

$$Q(\alpha) = \Psi \sum_{f=1}^{F} \sum_{m=1}^{M} \gamma_{m,f} \log \Big(\frac{1}{\sqrt{(2\pi)^N |\Sigma_m|}} \exp -\frac{1}{2} (\mathbf{A}_\alpha \boldsymbol{x}_f - \boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m{}^{-1} (\mathbf{A}_\alpha \boldsymbol{x}_f - \boldsymbol{\mu}_m) \Big) + \beta \log |\boldsymbol{A}_\alpha| \tag{20}$$

where, Ψ represents the scale factor for boosting likelihoods. The effect of the scale factor value "2" is shown in Figure 5b. The "optimal" scale factor is estimated empirically.

4) Using lower order features: VTLN is intended to transform the location of spectral peak positions, thus logically its estimation should be based upon only on the coarse spectral envelope. To illustrate this, reconstructed spectra from different MGCEP analysis orders are plotted in Figure 6. It can be noted that the cepstral order of 9 or more represents at least the first two formants. As the cepstral order increases beyond this, the finer details of the spectrum are better represented and the voicing information also



(a) Beta Prior Distribution (b) Scale factor

Fig. 5: Warping factors estimated from 39^{th} order features with a scale factor of 2 for the likelihoods and with a beta prior distribution. x-axis represents α values and y-axis represents the frequency of α values.

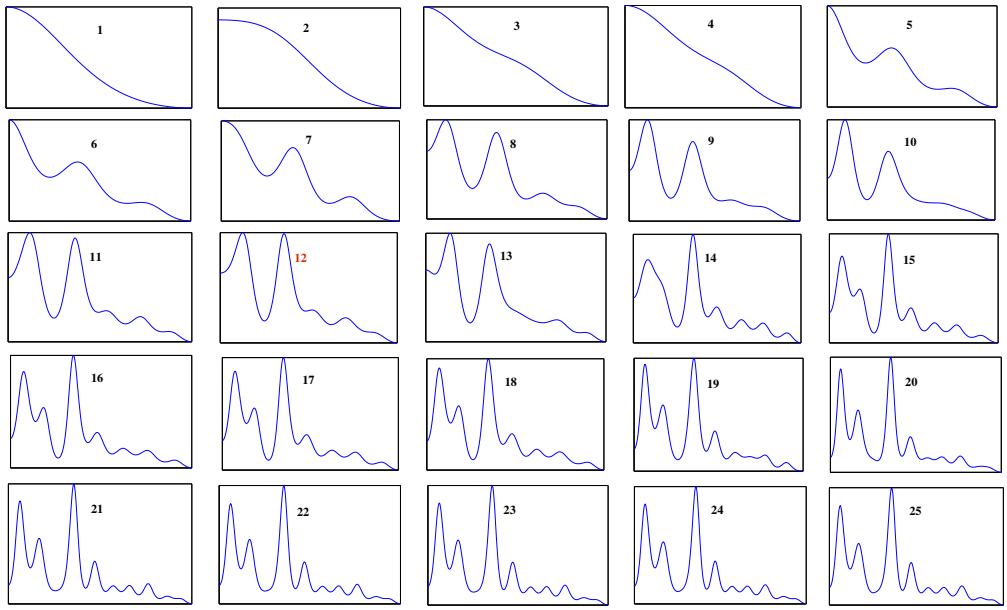


Fig. 6: Spectra reconstructed with cepstral features of the order of 1 to 25. The abscissa represents frequency and ordinate represents the spectral power.

creeps into the spectrum. We suppose that the spectral envelope as it relates to VTL is most accurately represented by an order around 12. This is consistent with use of feature order of 12 in the application of VTLN to ASR. In statistical parametric TTS, one means to overcome the problems posed by higher order cepstra in VTLN implementation. It is possible to estimate VTLN warping factors using only a subset of the cepstral features, specifically the first 13 (including C0). While this means that EM implementation of VTLN is no longer guaranteed to increase likelihood on the entire observation space, this work is more concerned with the implications of such an approach on subjective evaluations as this is of primary importance for TTS.

5) Approximation for inverse of the transformation matrix: Truncation of the cepstra and hence, truncation of the transformation matrix generates some inconsistencies in the calculation of the inverse

of the transformation matrix. This inconsistency can be attributed to the fact that the matrix A_α^{-1} does not represent an exact bilinear transformation matrix and hence, may not perform the desired warping.

Fortunately, the inverse of the bilinear transformation can be approximated as:

$$\mathbf{A}_\alpha^{-1} \approx \mathbf{A}_{-\alpha} \tag{21}$$

However, although equality is assumed in this paper, it should be noted that the expression is indeed an approximation (one just provides a more numerically stable way of computing the other). The matrices A_α^{-1} and $A_{-\alpha}$ are two possible approximations of the exact infinite transformation. While the former is more consistent with the estimation procedure, the latter is closer to an actual bilinear VTL warping, especially for lower order cepstral components.

This approximation can be justified by the fact that Equation 21 is exact in the infinite untruncated case and that it is used only during synthesis. Analysis of this approximation is outside the scope of this paper, but warrants further investigation. Further, this approximation does not have any effect on modelling or warping factor estimation. Only a determinant of the transformation matrix needs to be calculated during the warping factor estimation[1].

B. Subjective and perceptual factors

The concern in this paper is the use of probabilistic techniques for the application of VTLN to TTS; the ultimate goal is to find the best methods based on subjective evaluation. At times, this may require divergence from strict adherence to the theory, as in the case of using lower order cepstra to estimate VTLN warping factors. In short, subjective evaluation is the primary determinant of the success of the approaches presented in this paper, and as such we must understand perception of VTLN in statistical parametric TTS.

This section provides details on perceptual experiments that were carried out to establish a "subjective ground truth" for perception of VTL warping. These experiments also provide some insight to acoustic correlates of VTL (in particular pitch) and provide a basis for establishing a prior distribution of warping factors (as discussed in Section IV-A2). We also discuss the appropriate means to evaluate VTLN as speaker adaptation method, in particular, given its limitations with respect to reproducing specific voice characteristics.

1) Perceptual Impact of warping factors: In order to evaluate VTLN for speech synthesis we need to understand perception of VTL transformations, especially with respect to speaker similarity. In studying perception of VTL, we can also use data collected as the basis for constructing valid prior distributions

[1]Determinant is calculated as the product of eigen values especially for higher order features: $\log|\mathbf{A}| = \frac{1}{2}\sum_{i=1}^{N}\log\left(e_i e_i^*\right)$

for the warping parameter, α. This prior distribution can then have direct application in the MAP solution to the warping factor estimation.

VTL varies across speakers resulting in corresponding changes in the spectral peak positions. Conversely, warping the spectral frequencies of a recording 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 noted 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 original pitch of the speaker. These observations led to the design of a subjective evaluation to determine the perceived warping factors for a set of speakers. The values obtained from these evaluations are compared with the warping factors derived from the model.

Experiment design The HMM speech synthesis system (HTS) [25] was used to build average voice models using 39^{th} order cepstral features along with Δ and Δ^2 values of MGCEP features. Experiments were performed on the WSJCAM0 (British English) database with 92 speakers in the training set. The details of the synthesis system can be seen in Section V-B. Warped sentences were directly synthesized using the negative warping factors (instead of using default $\alpha = 0.42$), which is equivalent to applying the VTLN matrix $A_{(-\alpha)}$ to the untruncated cepstra.

Pitch and vocal tract length ideally should not be treated as independent parameters, although techniques have been proposed to implement VTLN based on pitch [28]. It is not in the scope of this work to implement such a system, but this factor is taken into consideration in the design of perceptual experiments, more specifically, the speakers selected in this experiment cover different combinations of pitch and VTL.

Experiments were performed only on the female speakers of the WSJCAM0 database. There were a total of 40 female speakers in the training set. A subset of 20 female speakers were selected in such a way that they cover the different possible combinations of pitch and VTL. 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 was equally divided into 3 sets: high, medium and low. Similarly, the range of α values derived using the average voice model for these speakers was also divided into 6 equally spaced groups. The warping factors were estimated using the EM approach described earlier using Jacobian normalization; issues of warping factor scaling, use of Jacobian normalization, etc. are not of concern since we are merely interested in dividing speakers into groups.

TABLE I: Frequency of female speakers with different combinations of vocal tract length and pitch

Pitch Vs "α" 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

The distribution of all (40) female speakers present in the training set according to this grouping is shown in Table I. A few significant observations can be made: There is a weak relationship between estimated warping factor and pitch with tendencies of low pitch speakers having lower warping and high pitch speakers to have higher warping. This suggests that most higher pitch voices are associated with females who have shorter VTL.

Natural pitch contours were extracted from the recorded speech of the selected speakers and speech was then resynthesized using the average voice models and the original pitch contours with six different warping factors in the range of 0 to 0.1. Listeners were asked to select the warping factor that synthesized the speech sounding closest to the target speaker. Thus, they selected a single option from the six different versions of the same utterance. Twenty-five listeners were asked to judge the speaker similarity in the speech files synthesized with different warping factors with reference to the natural speech from the speaker. This is repeated for 20 utterances each from a different speaker.

Observations and Discussion

The results of the subjective evaluations are shown in Figure 7. Each box represents a speaker in the evaluation set. It appears that listeners prefer some degree of warping (away from zero warping to give a clearly female voice characteristic). Extreme warping is also not preferred (too "childlike").

An analysis of the correlation between results from subjective evaluations and α derived from the HMM models is presented in Table II. The α values are derived from the models both with and without Jacobian normalization, which gives a different range for the warping factor distributions (as discussed in Section IV-A, the warping factors without Jacobian normalization have a higher range of α values). The table compares the values of mean, mode and median of the warping factors observed in the subjective evaluation. There is no statistically significant correlation between any values. The best correlation is seen between the median of the warping factors from subjective evaluation and those derived from the model without using Jacobian normalization.

Pitch does not show strong 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.

It is clear from these initial results that perception of VTL is a very difficult task to assess, but it is

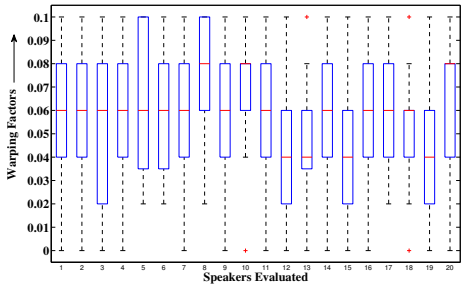


Fig. 7: Results from perceptual experiments. The ordinate represents the negative warping factor used to synthesize speech with female characteristics.

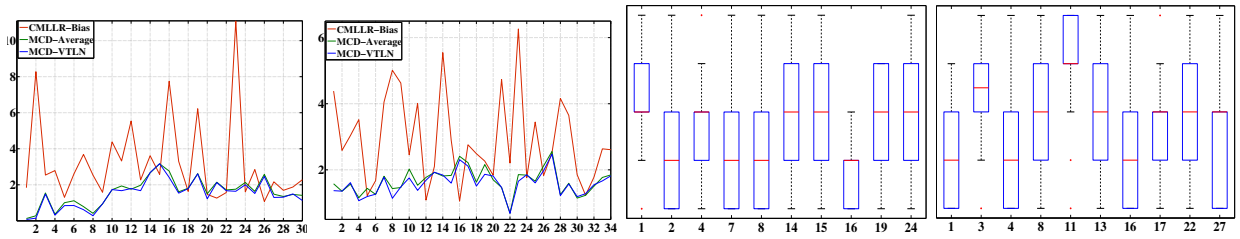
TABLE II: Correlation between model derived αs (with and without Jacobian normalization) and results of subjective evaluation. The mean, mode and median of the α values are derived from the results of subjective evaluations. Correlation between warping factors from all schemes and pitch is also presented.

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

evident that, on average, the preferred warping factors correspond to a distinctly female voice over one that remains close to the average voice. It is also clear that neither of the two baseline approaches (with and without Jacobian normalization) give impressive results. Thus, investigation of additional techniques as presented earlier has been further motivated. Towards this end, the experiment provides a good prior distribution of the warping factors for VTLN synthesis.

2) Evaluating VTLN performance: VTLN transforms only the spectral peaks, and hence very few speaker characteristics of the target speaker are introduced in the synthesized speech. If a target speaker has very different speaker characteristics when compared to the average voice model, owing to, for instance, accent, dialect, tempo or prosody, it can be difficult to evaluate the characteristics captured by VTLN. Furthermore, it is difficult to guarantee that listeners will evaluate speaker similarity based solely on the specified criteria, further skewing results.

The aim of this research is to evaluate VTLN as a spectral transformation ignoring other factors that cannot be captured by VTLN. To this end, experiments were performed to find speakers who can demonstrate characteristics captured by VTLN, and thus can be used to evaluate different VTLN adaptation techniques. Speakers were initially selected using two objective measures and then finally evaluated using subjective scores to find the best target speakers. Experiments were performed on two different English databases, WSJ0 (American English) and WSJCAM0 (British English). Average voice models were built separately using 83 and 92 speakers respectively in the two databases. The aim of this experiment is to find a male and female target speaker from the 30 and 34 selected test speakers of WSJ0 and WSJCAM0 respectively.



(a) WSJ0 Objective (b) WSJCAM0 Objective (c) WSJ0 Subjective (d) WSJCAM0 Subjective

Fig. 8: Objective and Subjective scores for WSJ0 and WSJCAM0 databases. For objective scores, ordinate represents the value for MCD or magnitude of the bias term. For subjective scores, ordinate represents the DMOS score between 1 and 5. Abscissa in all figures represents index of the test speakers

It is postulated that adaptation works better on speakers closer to the average voice model [29]. The objective scores are chosen so as to find the speakers closer to average voice. The first objective score is based on the bias term in MLLT adaptation techniques, which should ideally represent the displacement of the speaker parameters from the average voice model. CMLLR transforms were generated for the target speakers using the average voice models, and the magnitude of the bias vector was calculated for each speaker. The second objective score is based on the mel-cepstral distortion (MCD) value which is the Euclidean distance between cepstra.

Speech parameters were generated from the average voice using VTLN adaptation for each of the test speakers, and MCD scores were calculated between the synthesized parameters and the parameters extracted from the natural speech of the target speaker. Both MCD and CMLLR-bias values for the speakers from two databases are plotted in Figures 8(a-b). The figure also shows the MCD scores between average voice and the natural speech of the speaker (denoted as "MCD-Average") which displays the same trend as the score between VTLN synthesized voice and natural speech (denoted as "MCD-VTLN"). It is noted that there is not much correlation between the two objective scores and it is not feasible to select target speakers only based on these scores.

Subjective evaluations were designed in order to find speakers that are rated better with VTLN adaptation. Since it is onerous to perform evaluations on all the test speakers, 10 test speakers were selected from each database. 5 speakers with low MCD scores and 5 with high MCD scores were considered, taking account of gender balance. Listeners were asked to rate the synthesized speech based on speaker similarity on a 5 point scale with 5 being "Sounds like exactly the same speaker" and 1 being "Sounds like an entirely different speaker".

Results plotted as mean opinion scores (MOS) are presented in Figures 8(c-d). The same speaker indices are used in both subjective and objective plots. The results from these evaluations are summarized in Table III. Speakers are marked as "1" (preferred) and "0" (not preferred) accordingly in the table for

TABLE III: Summary of speaker selection experiments for WSJ0 and WSJCAM0. 'M/F' stands for Male/Female speaker. Speakers are classified as having high or low scores. Preferred speakers should ideally have "low" objective scores (CMLLR-Bias & MCD) and "high" subjective scores (MOS). For the sake of improved readability the scores are marked "1" (preferred) and "0" (not preferred) accordingly in the table.

Speaker#	M/F	MOS	MCD	Bias
1	**M**	**1**	**1**	**1**
2	F	0	1	0
4	M	1	1	1
7	F	0	1	1
8	F	0	1	1
14	M	1	0	1
15	M	1	0	1
16	F	0	0	0
19	M	1	0	0
24	**F**	**1**	**0**	**1**

Speaker#	M/F	MOS	MCD	Bias
1	F	0	1	0
3	F	1	0	0
4	M	0	1	0
8	M	1	1	0
11	**F**	**1**	**1**	**0**
13	M	1	0	1
16	F	0	0	1
17	F	1	0	1
22	**M**	**1**	**1**	**1**
27	M	1	0	1

each score. It is noted that the subjective results do not have much correlation with any of the objective results. Both scores agree for only a single speaker in both databases. A male and female test speaker with the best subjective score is selected for each database. Conflicts are resolved by a voting scheme between the scores giving least priority to CMLLR-bias and highest priority to subjective results. The next section presents the details of the experiments performed to evaluate VTLN techniques and the corresponding results.

V. Evaluations with VTLN

This section presents experiments carried out to evaluate the performance of different VTLN approaches in statistical parametric speech synthesis framework. More specifically, the baseline bilinear transform-based VTLN using EM learning criterion, as presented in Section III, is compared against several variants based on the methods presented in Section IV. The goal of this work is to find the most effective VTLN approach for statistical parametric speech synthesis and, by way of this, explain some of the differences observed in comparing past literature on VTLN to our own studies. In comparing different VTLN estimation techniques for synthesis we are not aiming to surpass the performance of more powerful adaptation approaches such as CMLLR. Instead, our longer-term goal is to combine VTLN with such adaptation approaches, as will be discussed further in the conclusions.

A. Techniques evaluated

As explained in section IV, there are various techniques to overcome the challenges in estimating warping factors for TTS. The approaches discussed were primarily introduced to address higher TTS feature dimensionality and included:

TABLE IV: Techniques to be evaluated for VTLN warping factor estimation

Name	Technique
T1	Jacobian Normalization
T2	No Jacobian Normalization
T3	Jacobian Normalization with scaled prior
T4	Jacobian Normalization with scaled likelihood
T5	Combination of T3 and T4
T6	Using feature blocksize 13 & Jacobian Normalization
T7	Using feature blocksize 13 & No Jacobian Normalization

a) Omission of Jacobian normalization: as has previously been done in ASR studies;

b) Prior distribution of warping factors: to provide a better estimate in test conditions when there is little adaptation data;

c) Likelihood scaling: applied to acoustic scores (while omitting scaling for the Jacobian); and

d) Lower order features: for the estimation of warping factor, thereby ignoring high order cepstra in EM auxiliary function;

All the techniques evaluated in this section are summarized in Table IV. The same nomenclature as in the table is used when labeling the results. The scale factors for prior (value of '10') and likelihood (value of '3') are estimated empirically in cases (b) and (c) above (for T3, T4 and T5 in Table IV). Evaluations are also to be performed for both single and multiple (regression class-based) transform VTLN. It would have been ideal to estimate different prior distributions for each class in a multi-class VTLN system. Since this is out of the current scope of this research, same prior distributions are used for different regression classes in the multi-class VTLN system. This is one of the possible future directions for this work.

TABLE V: MCD (in dB) for VTLN synthesis for WSJ0 and WSJCAM0 with 10 test speakers.AV represents average voice.

	Global	Multiple
AV	7.616	-
T1	7.518	7.517
T2	7.469	7.464
T3	7.439	7.500
T4	7.464	7.469
T5	7.441	7.465
T6	7.456	7.433
T7	7.527	7.507

	Global	Multiple
AV	6.107	-
T1	5.994	5.982
T2	6.000	5.980
T3	5.973	5.964
T4	6.010	6.018
T5	5.988	5.946
T6	5.986	5.973
T7	5.977	5.975

B. Experimental Setup

The HMM speech synthesis system (HTS), more specifically the system scripts for the HTS-2007 submission to the Blizzard Challenge [25] provided the basis for training and generating the statistical parameters for speech synthesis. HTS models spectrum, pitch (log f0), band-energy and duration in the unified framework of hidden semi-Markov models (HSMMs). Features extracted were 39th-order mel-cepstra[2] derived from STRAIGHT spectrum, log F0, five-band aperiodicity, and their delta and delta-delta coefficients, extracted from 16kHz recordings with a window shift of 5ms. The STRAIGHT vocoder was used to synthesize speech from the parameters generated using HTS.

Two English average voice synthesis model sets were trained on the WSJ0 and WSJCAM0 corpora using an HMM five-state left-to-right topology. Models were trained using CMLLR-based speaker adaptive training (CMLLR-SAT), while all feature extraction was performed with $\alpha = \alpha_{\mathrm{M}} = 0.42$, which was applied to the untruncated cepstrum. We undertook this approach in order to facilitate the evaluation of different VTLN estimation techniques using a common canonical model. For synthesis, the VTLN warping factor, α_{V}, was estimated as a model transform using EM approach as described in this paper. However, for synthesis, the negative of α_{V} was used as a model transform (implemented by multiplying $\mathsf{A}_{-\alpha}$ matrix with the 40 dimensional model parameters and synthesizing features with the transformed models).

For objective scores, different VTLN warping factor estimation techniques were evaluated for ten target speakers from each database (5 male and 5 female). Only a single sentence from each target speaker was used as the adaptation data. The objective measures are based on mel-cepstral distortion (MCD). MCD is the Euclidean distance of synthesized cepstra from that of the values derived from the natural speech.

The subjective performance was evaluated based on naturalness and speaker similarity using MOS and DMOS scoring respectively. The synthesized utterances were rated on a 5-point scale, 5 being "completely natural" or "sounds like the exact target speaker" and 1 being "completely unnatural" or "sounds like a totally different speaker". For subjective testing, one male and one female test speaker from each database were selected based on experiments presented in IV-B2. The subjective evaluations were performed using Amazon Mechanical Turk[3] online evaluation setup. The listeners were paid for their service.

C. Results and Discussion

This section presents the results of the evaluations and discusses some of the conclusions that can be drawn from these results.

[2]SPTK's `mcep` function was used, which is equivalent to mel-generalised ceptrum with $\gamma = 0$ while it does not apply the UELS optimisation criterion.

[3]https://www.mturk.com

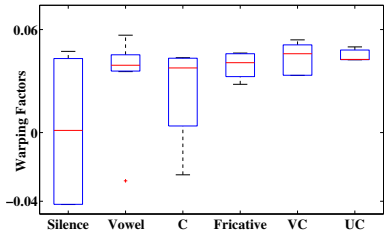
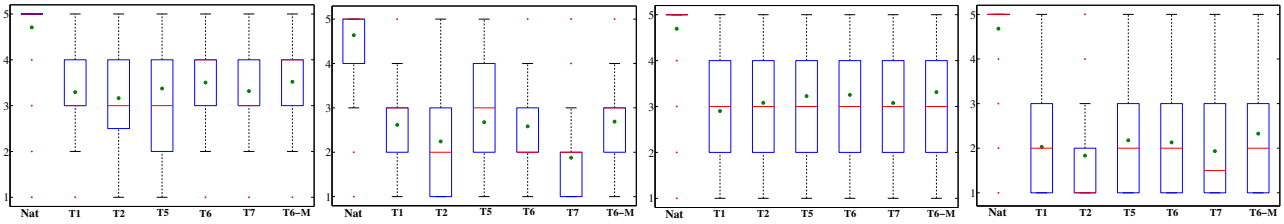


Fig. 9: Distribution of α for different phoneme classes ('C' = all consonants, 'VC' = voiced consonants, 'UC' = unvoiced consonants) for a specific male speaker. The global VTLN warping factor in this case is "0.0467". Note that C $\neq$ VC $\cup$ UC; see the text.



(a) WSJ0 Database (b) WSJCAM0 Database (c) WSJ0 Database (d) WSJCAM0 Database

Fig. 10: Subjective Scores for Naturalness (a-b) and Speaker Similarity (c-d) for two speakers in each database. Nat represents Vocoded speech and all VTLN systems use only a single parameter except for T6-M system which is the multiple parameter version of the T6 system. Abscissa represents different systems and ordinate represents the MOS scores.

1) Analysis of Warping Factors: Some analysis of warping factors obtained by the proposed approach for multi-class VTLN is made first. The distribution of α for different phoneme classes for a male speaker is shown in the Figure 9. The values were derived using Jacobian normalization with scaled likelihood and prior. Using this method results in warping factors that are slightly biased towards the prior for all classes which explains the high warping factors for some consonant classes. A clearer difference would be observed between classes in the case where no prior is used. 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. Consonants, in general, display warping factors tending to lower than average values. Whilst the consonants class represents all consonants, the voiced and unvoiced categories of consonants shown in the figure are the small subsets that had *only* those labels, rather than combinations of those and other labels such pulmonic or plosive. These two pure classes show somewhat opposite trends to one another.

2) Objective Results: The objective evaluations based on MCD are performed for all the 10 speakers selected for the experiments in section IV-B2. All the techniques proposed in this paper are evaluated objectively. The abbreviations for the systems are presented in Table IV and MCD results in Table V. It was observed that a few speakers/systems give extreme warping factors which result in huge MCD scores which could be fixed by using $A_{-\alpha}$ instead of A_{α}^{-1} during synthesis.

TABLE VI: Wilcoxon signed rank test for significance of 1% for Naturalness for WSJ0 and WSJCAM0.

	Nat	T1	T2	T5	T6	T7	T6-M
Nat	-	1	1	1	1	1	1
T1	1	-	0	0	1	0	1
T2	1	0	-	1	1	0	1
T5	1	0	1	-	0	0	0
T6	1	1	1	0	-	1	0
T7	1	0	0	0	1	-	1
T6-M	1	1	1	0	0	1	-

	Nat	T1	T2	T5	T6	T7	T6-M
Nat	-	1	1	1	1	1	1
T1	1	-	1	0	0	1	0
T2	1	1	-	1	1	1	1
T5	1	0	1	-	0	1	0
T6	1	0	1	0	-	1	0
T7	1	1	1	1	1	-	1
T6-M	1	0	1	0	0	1	-

TABLE VII: Wilcoxon signed rank test for significance of 1% for Speaker similarity for WSJ0 and WSJCAM0.

	Nat	T1	T2	T5	T6	T7	T6-M
Nat	-	1	1	1	1	1	1
T1	1	-	0	1	1	0	1
T2	1	0	-	0	0	0	1
T5	1	1	0	-	0	0	0
T6	1	1	0	0	-	0	0
T7	1	0	0	0	0	-	1
T6-M	1	1	1	0	0	1	-

	Nat	T1	T2	T5	T6	T7	T6-M
Nat	-	1	1	1	1	1	1
T1	1	-	1	0	0	0	1
T2	1	1	-	1	1	0	1
T5	1	0	1	-	0	1	0
T6	1	0	1	0	-	1	0
T7	1	0	0	1	1	-	1
T6-M	1	1	1	0	0	1	-

3) Subjective Results: In order to simplify the listening tests, not all techniques are subjectively evaluated. Techniques to be evaluated subjectively are selected based on the objective results. Except for the T6-M system, all systems in subjective evaluation use the global transform based VTLN. The T6-M system represents the multiple transform case for the T6 system. From the observations based on earlier research [6] and the objective results, T3 and T4 systems are not very different from their combination T5 system. Hence, these systems are omitted from subjective evaluation. 75 listeners participated in the evaluation and listened to 124 sentences in total. The results for naturalness and speaker similarity are plotted in Figure 10. Table VI and Table VII shows if the systems are significantly different based on the Wilcoxon signed rank test (with significance of 1%). A few systems are shown to be significantly different from each other. Listeners do not prefer systems that did not use Jacobian normalization. The systems that use lower order features to estimate the VTLN warping factors are prefered in general (both while using single and multiple VTLN parameters).

Both objective and subjective results for systems not using Jacobian normalization (systems T2 and T7) are worse when compared to other systems. This can be attributed to the fact that not using Jacobian normalization not only gives higher values for warping factors, but also sometimes gives extreme values (as shown in Figure 3d). This extreme warping is not preferable for the test speakers when evaluating speaker similarity or naturalness and also results in the huge MCD values for these speakers unless the inverse of the α matrix is calculated carefully. This same phenomena was observed when using the likelihood scaling (system T4) some test speakers. It can be concluded from these results that Jacobian normalization is an important factor that should not be avoided for estimating VTLN warping factors,

especially when using a large number of test speakers, who might give some extreme results when not using Jacobian.

However, at the same time, it can be observed that using Jacobian normalization directly with higher order features (system T1) is not very preferable especially for speaker similarity or MCD scores. Using Jacobian normalization with higher order features over compensates the variations in the warping factors and reduces the spread of warping factors. This in turn results in less discrimination among different test speakers and further reduces the speaker similarity. These observations motivate the use a technique that can overcome the problem of feature dimensionality such as using the lower order features for estimating the warping factor (systems T6 or T7), which is a novel technique presented in this paper.

From the above experiments, it can be concluded that the best method to estimate VTLN warping factors for statistical parametric speech synthesis is using lower order features with Jacobian normalization (system T6). This technique gives stable objective results for all speakers and comparatively better subjective evaluation performance. Though there was minor improvements in the objective scores, the multiple transform case (using regression class trees) for this system could not display any statistically significant improvement in the performance over the global transform based VTLN adaptation. The MAP based VTLN parameter estimation using prior distributions (system T5) also gives almost the same values for the warping factors and similar performance as using the lower order features, T6 system.

VI. Conclusion

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 applies different warping factors to different state distributions. 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 are only limited characteristics that this parameter can capture. Hence, in order to obtain improvements in adaptation when more adaptation data is available, VTLN may need to be combined with other adaptation methods such as CMLLR. In order to combine VLTN with CMLLR, the author's current research focusses on implementing VTLN as a prior to CMLLR transform as in constrained structural maximum a posteriori linear regression (CSMAPLR).

Acknowledgment

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). The authors would like to thank the reviewers for their valuable contribution in improving the paper.

REFERENCES

[1] A. W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *Proc. of ICASSP*, Hawaii, USA, 2007, pp. 1229–1232.

[2] 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, no. 1, pp. 66–83, Jan. 2009.

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

[4] 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.

[5] L. Saheer, P. N. Garner, J. Dines, and H. Liang, "VTLN adaptation for statistical speech synthesis," in *Proc. of ICASSP*, Mar. 2010, pp. 4838–4841.

[6] L. Saheer, J. Dines, P. N. Garner, and H. Liang, "Implementation of VTLN for statistical speech synthesis," in *Proc. of the 7th ISCA Speech Synthesis Workshop*, Kyoto, Japan, Sep. 2010, pp. 224–229.

[7] L. Saheer, P. N. Garner, and J. Dines, "Study of Jacobian normalization for VTLN," *Idiap-RR-25-2010*, 2010.

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

[9] S. Umesh, A. Zolnay, and H. Ney, "Implementing frequency warping and VTLN through linear transformation of conventional MFCC," in *Proc. of Interspeech*, Lisbon, Portugal, 2005, pp. 269–271.

[10] 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.

[11] 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.

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

[13] L. F. Uebel and P. C. Woodland, "An investigation into vocal tract length normalisation," in *Proc. of the European Conference on Speech Communication and Technology*, 1999, pp. 2527–2530.

[14] D. Sundermann, "Text-independent voice converision," Ph.D. dissertation, Bundeswehr University Munich, Munich, Germany, 2008.

[15] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis – A unified approach to speech spectral estimation," in *Proc. of ICSLP*, vol. 3, Sep. 1994, pp. 1043–1046.

[16] 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.

[17] 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 *Proc. of Interspeech*, Brisbane, Australia, 2008, pp. 1713–1716.

[18] D. Oppenheim, A.V. Johnson, "Discrete representation of signals," *Proc. of IEEE*, vol. 60, pp. 681–691, 1972.

[19] A. Sankar and C.-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 3, pp. 190–202, May 1996.

[20] T. Emori and K. Shinoda, "Rapid vocal tract length normalization using maximum likelihood estimation," in *Proc. of Eurospeech*, 2001, pp. 1649–1652.

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

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

[23] S. Molau, S. Kanthak, and H. Ney, "Efficient vocal tract normalization in ASR," in *Proc. of ESSV*, Cottbus, Germany, 2000.

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

[25] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 6, pp. 1208–1230, Aug. 2009.

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

[27] 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 *Proc. of Interspeech*, Brighton, UK, 2009, pp. 584–587.

[28] A. Faria and D. Gelbart, "Efficient pitch-based estimation of VTLN warp factors," in *Proc. of Interspeech*, Sep. 2005, pp. 213–216.

[29] J. Yamagishi, O. Watts, S. King, and B. Usabaev, "Roles of the average voice in speaker-adaptive HMM-based speech synthesis," in *Proc. of Interspeech*, Sep. 2010, pp. 418–421.

APPENDIX

Derivation for Jacobian Normalization

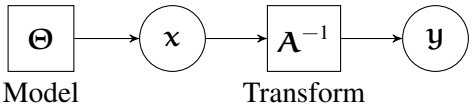


Fig. 11: Generative model for vocal tract "warping"

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

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}}\, \mathrm{p}_{\alpha}\left(\alpha \mid \mathbf{y}, \boldsymbol{\Theta}\right) \propto \mathrm{p}_{\mathbf{y}}\left(\mathbf{y} \mid \alpha, \boldsymbol{\Theta}\right) \mathrm{p}_{\alpha}\left(\alpha \mid \boldsymbol{\Theta}\right). \tag{22}$$

To evaluate the first term on the RHS of equation 22, notice that the model generates $\mathbf{x}$ rather than $\mathbf{y}$, so it needs a change of variable $\mathbf{y} \rightarrow \mathbf{x}$. The Jacobian determinant for the change of variable is,

$$\mathrm{J} = |\mathbf{A}|\,, \tag{23}$$

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

$$\mathrm{p}_{\mathbf{y}}\left(\mathbf{y} \mid \alpha, \boldsymbol{\Theta}\right) = |\mathbf{A}|\,\mathrm{p}_{\mathbf{x}}\left(\mathbf{A}\mathbf{y} \mid \alpha, \boldsymbol{\Theta}\right). \tag{24}$$

The second term on the RHS of equation 22 is a prior on α. Notice that α is actually independent of the model, $\boldsymbol{\Theta}$, so it could be written unconditional. However, α is posterior to the training data, $\mathbf{D}$, that was used to train $\boldsymbol{\Theta}$. So, equation 22 can be evaluated as

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}}\, |\mathbf{A}|\,\mathrm{p}_{\mathbf{x}}\left(\mathbf{A}\mathbf{y} \mid \alpha, \boldsymbol{\Theta}\right)\mathrm{p}_{\alpha}\left(\alpha \mid \mathbf{D}\right). \tag{25}$$

Notice that the prior is not normally considered.