Combining Vocal Tract Length Normalization with Hierarchical Linear Transformations

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

***Abstract*—Recent research has demonstrated the effectiveness of vocal tract length normalization (VTLN) as a rapid adaptation technique for statistical parametric speech synthesis. VTLN produces speech with naturalness preferable to that of MLLR-based adaptation techniques, being much closer in quality to that generated by the original average voice model. However, with only a single parameter, VTLN captures very few speaker specific characteristics when compared to linear transform based adaptation techniques. This paper shows that the merits of VTLN can be combined with those of linear transform based adaptation in a hierarchical Bayesian framework, where VTLN is used as the prior information. A novel technique for propagating the gender and age information captured by the VTLN transform into constrained structural maximum a posteriori linear regression (CSMAPLR) adaptation is presented. This paper also compares this proposed technique to other combination techniques. Experiments are performed on both matched and mismatched training and test conditions, including gender, age, and recording environments. Text-to-speech (TTS) synthesis experiments show that the resulting transformation produces improved speech quality with better naturalness and intelligibility (similar to VTLN transformation) when compared to the CSMAPLR transformation, especially when the quantity of adaptation data is very limited. With more parameters to capture speaker characteristics, the proposed method performs better in speaker similarity compared to VTLN in mis-matched conditions. Hence, the proposed combination combines the quality and intelligibility of VTLN with the speaker similarity of CSMAPLR especially in the mismatched train and test conditions. Experiments are also performed using the automatic speech recognition (ASR) system in a unified framework as that of synthesis. This is to prove that the techniques developed for TTS can be plugged into ASR in order to improve the performance.**

***Index Terms*—Statistical parametric speech synthesis, hidden Markov models, speaker adaptation, vocal tract length normalization, constrained structural maximum a posteriori linear regression**

I. Introduction

The ability to transform voice identity in text-to-speech synthesis (TTS) has been an important area of research with applications in the 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, which can help overcome the language barrier, especially on a mobile device. It is crucial to this kind of application that the speaker characteristics are introduced into the output speech from the very first utterance spoken by a speaker. Hence, speaker characteristics need to be estimated from very little adaptation data.

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 [2] previously developed for automatic speech recognition (ASR). 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. These techniques require adaptation data including tens of utterances for reasonable adaptation performance. Rapid adaptation techniques like vocal tract length normalization (VTLN) have also been successfully applied to statistical parametric speech synthesis [3, 4, 5]. By contrast, this technique requires very little adaptation data as it estimates only a single parameter. This approach preserves the naturalness of the average voice, albeit capturing very few speaker characteristics. It follows that combining the linear transform based adaptation techniques with VTLN could result in improved naturalness of synthesized speech whilst also being effective at capturing the speaker characteristics. This provides a means to rapidly adapt synthesized speech with a balanced trade-off between naturalness and speaker similarity.

VTLN is a widely used speaker normalization technique in ASR [6, 7]. It is inspired from the 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 [8]. The formant frequency positions are inversely proportional to VTL, and hence can vary around 25% [9]. Although implementation details differ, VTLN is generally characterized by a single parameter that warps the spectra 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 also estimate the speaker characteristics of a target speaker, and hence transform the average voice into the speech of the target speaker. Initial investigations of VTLN for statistical parametric speech synthesis were performed by Saheer et. al. [10].

Breslin et al. [11] showed that VTLN can be combined with constrained MLLR (CMLLR) for rapid adaptation in

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 HASLER funded V-FAST project, EPSRC grants EP/I031022/1 (NST), and EP/J002526/1 (CAF). The current research is funded by D-Box project.

L. Saheer, P. N. Garner and J. Dines are with Idiap Research Institute, Switzerland. J. Yamagishi is with Centre for Speech Technology Research, University of Edinburgh, U.K. and with National Institute of Informatics, Japan. L. Saheer is also affiliated with Ecole Polytechnique Fédérale de Lausanne, Switzerland. e-mail: (lsaheer,dines,pgarner)@idiap.ch. jyamagis@inf.ed.ac.uk

Manuscript received March 26, 2012.

ASR. In that work, a count smoothing framework is used to incorporate the prior information. In this paper, we focus on structural maximum a posteriori (SMAP) based adaptation techniques that use prior information for transform estimation in a hierarchical way [12] – The SMAP technique uses a family of elliptically symmetric distributions including the matrix variate normal prior density as a prior distribution [13] and uses a tree structure to propagate this prior to different classes of transforms. Yamagishi et. al. [2] showed that due to the presence of hierarchial prior, constrained SMAP linear regression (CSMAPLR) is a more robust adaptation framework when compared to CMLLR in statistical parametric speech synthesis.

There are a number of potential ways of combining VTLN with the CSMAPLR based linear transformation framework, including as a cascade of linear transforms in a similar way to ASR. In this paper we explore a more effective and mathematically consistent way. More specifically, we treat the VTLN transform as the Bayesian prior for CSMAPLR and derive a hyper-parameter for the CSMAPLR adaptation at the root node from a VTLN transform. The structural framework of the SMAP criterion helps propagate the prior information affected by the VTLN transform into the various levels of the regression tree seamlessly and effectively. Using the VTLN matrix as the initial prior information for the CSMAPLR transform at the root node could result in better propagation of gender and age characteristics and hence improved speaker adaptation even when very little data is available.

Both speech synthesis and recognition experiments are performed in a unified framework representing the most favourable scenario of a speech-to-speech translation system. These experiments do not represent the state-of-the-art results in ASR, rather prove the point that similar techniques can be adopted and can prove advantageous to both HMM-based speech synthesis and recognition. The experiments are performed on matched and mismatched train and test conditions. Matched conditions include speakers of the similar gender, age, or speech recorded in similar environmental conditions for training and testing. Three forms of mismatched conditions evaluated in this paper include

1) the gender mismatch where gender dependent male or female models are used to test speakers of other gender,
2) the age mismatch where average voice models trained on adult speech were adapted into child voice, and
3) the recording environment mismatch where speech in different noise conditions are tested on models trained with clean speech.

Since it is known that VTLN performs better in the extreme mismatch conditions, the combinations of VTLN and CSMAPLR are also expected to give improvements in these scenarios.

The paper is organised as follows: Details on the VTLN and CSMAPLR based linear transformations are presented in section II. The proposed technique and several different ways to combine VTLN with CSMAPLR are presented in section III, followed by the experiments in the matched train and test conditions in section IV. The results for mismatched conditions are presented in section V. Finally, observations and conclusions are given in section VI.

II. Overview

A. VTLN and CMLLR

The main components involved in VTLN are a warping function, a warping factor and an optimization criterion. Typically, the warping function has only a single variable α as the warping factor, which is representative of the ratio of the VTL of a speaker to an average VTL.

In ASR, where a mel or bark spaced filter bank is used, the warping function tends to be linear or piecewise-linear, and is normally applied directly to the filter-bank. By contrast, feature extraction for TTS systems tends not to use a filter-bank analysis as it renders signal reconstruction difficult. Rather, the feature commonly used in TTS is the mel-generalized cepstrum (MGCEP) [14], which makes use of a bilinear transform to achieve a frequency warp[1]. Since MGCEP already includes a bilinear transform, a bilinear transform-based VTLN proposed by Pitz and Ney [15] can be implemented as a zero-overhead modification of the MGCEP representation.

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:

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

where $z^{-1} = e^{-j\omega}$, $\tilde{z}^{-1} = e^{-j\tilde{\omega}}$, and α is the warping factor. We define the m-th mel-cepstral coefficient, that is, frequency warped cepstrum, $\tilde{c}_m$ in MGCEP as

$$\tilde{c}_m = \frac{1}{2\pi j} \oint_C \log X(\tilde{z})\, \tilde{z}^{m-1} d\tilde{z} \tag{2}$$

$$\log X(\tilde{z}) = \sum_{m=-\infty}^{\infty} \tilde{c}_m\, \tilde{z}^{-m} \tag{3}$$

Since the frequency warping is $X(\tilde{z}) = X(z)$, we have a linear transformation in the cepstral domain c_k:

$$\tilde{c}_m = \sum_{k=-\infty}^{\infty} \frac{1}{2\pi j} \oint_C \tilde{z}^{-k} z^{m-1} d\tilde{z}\, c_k \tag{4}$$

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

where $A_{mk}(\alpha)$ is the m-th row k-th column element of the warping matrix $\boldsymbol{A}_\alpha$ consisting of the warping factor α and the Cauchy integral formula yields [15]:

$$A_{mk}(\alpha) = \frac{1}{2\pi j} \oint_C \tilde{z}^{-k} z^{m-1} d\tilde{z} \tag{6}$$

$$= \frac{1}{2\pi j} \oint_C \left(\frac{z - \alpha}{1 - \alpha z} \right)^{-k} z^{m-1}\, d\tilde{z} \tag{7}$$

$$= \frac{1}{(k-1)!} \sum_{n=\max(0,k-m)}^{k} \binom{k}{n} \times \frac{(m+n-1)!}{(m+n-k)!} (-1)^n \alpha^{2n+m-k}. \tag{8}$$

[1]Spectral analysis in MGCEP also uses a generalized logarithmic function, which has the effect of varying the analysis between an all-pole and a cepstral model, according to a second parameter.

We may represent the transformation in the vector form $\boldsymbol{x}_\alpha = \boldsymbol{A}_\alpha \boldsymbol{x}$, where $\boldsymbol{x}_\alpha = (\tilde{c}_1, \cdots, \tilde{c}_M)^\top$ and $\boldsymbol{x} = (c_1, \cdots, c_K)^\top$ if we truncate the original and warped mel-cepstral coefficients at K-th and M-th dimensions. The transform may also be directly applied to the "dynamic" features of the cepstra, where the transformation matrix is block diagonal with repeating $\boldsymbol{A}_\alpha$ matrix.

The maximum likelihood criterion can be adopted for the optimisation of the warping factor α [9]:

$$\widehat{\alpha}_s = \underset{\alpha_s}{\operatorname{argmax}} P(\boldsymbol{x}_{1,\alpha_s}, \boldsymbol{x}_{2,\alpha_s}, \ldots, \boldsymbol{x}_{T,\alpha_s} \mid \Theta, \alpha_s, w_s) \quad (9)$$

where $\boldsymbol{x}_{t,\alpha_s}$ represents features at time t, warped with the warping factor α_s for speaker s; T is the total number of frames; Θ represents average voice models, w_s represents the word sequence corresponding to features and $\widehat{\alpha}_s$ represents the optimal warping factor for speaker s.

VTLN can also be implemented as an equivalent feature-space MLLT using $\boldsymbol{A}_\alpha$; such representation enables use of the EM algorithm for finding optimal warping factors. The main advantage of using the EM algorithm over, say, a grid search is that the resulting warping factor estimation has finer granularity of α values, and efficient implementation in time and space. The EM algorithm can be embedded into HMM training utilizing the same sufficient statistics as CMLLR [7], which transforms the spectral features as follows

$$\tilde{\boldsymbol{x}}_t = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{b} = \boldsymbol{W}\boldsymbol{\xi}_t. \quad (10)$$

where $\boldsymbol{\xi}_t = [\boldsymbol{x}_t^\top, 1]^\top$, and $\boldsymbol{W} = [\boldsymbol{A}, \boldsymbol{b}]$. Note that, the matrix $\boldsymbol{A}$ and bias vector $\boldsymbol{b}$ of the CMLLR transform are far less constrained than those for VTLN. The VTLN transform is known to represent the changes due to the differences in the length of the vocal tract among individuals. The maximum change for the vocal tract length results in a 25% change in the spectral peaks ranging from a factor of -0.1 to +0.1. This restricts the VTLN transformation also to take values within this range. Hence, VTLN is a constrained transformation compared to CMLLR. Similar to CMLLR, VTLN represents a transformation of the spectral parameters for a speaker, but, based on his/her physical characteristics. More specifically, the number of free parameters in VTLN transformation is one, while, in CMLLR is a complete transformation matrix.

B. CSMAPLR

CSMAPLR is a robust framework to estimate the CMLLR transforms $\boldsymbol{W}$ based on the SMAP criterion [12]:

$$\widehat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\operatorname{argmax}} P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \mid \Theta, \boldsymbol{W}, w_s)\, P(\boldsymbol{W}) \quad (11)$$

where $\boldsymbol{W}$ refers to the set of CMLLR transforms. $P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \mid \Theta, \boldsymbol{W}, w_s)$ is a likelihood function for $\boldsymbol{W}$ and $P(\boldsymbol{W})$ is a prior distribution of the transform $\boldsymbol{W}$. Matrix variate normal distributions are used as the prior distribution $P(\boldsymbol{W})$:

$$P(\boldsymbol{W}) \propto |\boldsymbol{\Omega}|^{-\frac{L+1}{2}} |\boldsymbol{\Psi}|^{-\frac{L}{2}} \exp\left[-\frac{1}{2}\mathrm{tr}(\boldsymbol{W}-\boldsymbol{H})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{W}-\boldsymbol{H})\boldsymbol{\Psi}^{-1}\right] \quad (12)$$

where $\boldsymbol{\Omega} \in \mathbb{R}^{L\times L}$, $\boldsymbol{\Psi} \in \mathbb{R}^{(L+1)\times(L+1)}$ and $\boldsymbol{H} \in \mathbb{R}^{L\times(L+1)}$ are the hyperparameters of the prior distribution.

In the CSMAPLR estimation, the hyperparameter $\boldsymbol{\Psi}$ is fixed to the identity matrix and $\boldsymbol{\Omega}$ to a scaled identity matrix, $\boldsymbol{\Omega} = \tau \boldsymbol{I}_L$. τ is a positive scalar that controls the scale factor for the prior propagation and $\boldsymbol{I}_L$ is $L \times L$.

In the SMAP criterion, the tree structures of the distributions called "regression class tree" effectively control these hyperparameters. First, at the root node of the regression class tree, a transform $\boldsymbol{W}_1$ is estimated using all available adaptation data and the ML criterion. A new transform at a child node 2 represented as $\boldsymbol{W}_2$ is then MAP estimated using the corresponding adaptation data and the transform $\boldsymbol{W}_1$ as a hyperparameter $\boldsymbol{H}$ of the prior distribution, that is, $\boldsymbol{H} = \boldsymbol{W}_1 = [\boldsymbol{A}_1, \boldsymbol{b}_1]$. Likewise, a new transform $\boldsymbol{W}_3$ at a grandchild node 3 is further MAP estimated using the corresponding adaptation data and the transform $\boldsymbol{W}_2$ as a hyperparameter (i.e. $\boldsymbol{H} = \boldsymbol{W}_2 = [\boldsymbol{A}_2, \boldsymbol{b}_2]$). This process is continued recursively from the root node to all the leaf nodes of the tree structure.

The re-estimation formula based on the Baum-Welch algorithm for the transformation matrix is given by [16]:

$$\hat{\boldsymbol{w}}_l = (\kappa \boldsymbol{p}_l + \boldsymbol{k}_l){\boldsymbol{G}_l}^{-1} \quad (13)$$

where $\hat{\boldsymbol{w}}_l$ represents the l-th row of the transform $\boldsymbol{W}$, $\boldsymbol{p}_l = [0, \boldsymbol{c}_l]$, and $\boldsymbol{c}_l$ is the l-th cofactor row vector of the transform $\boldsymbol{W}$. The value κ satisfies the quadratic equation:

$$\kappa^2 \boldsymbol{p}_l {\boldsymbol{G}_l}^{-1} \boldsymbol{p}_l^\top + \kappa \boldsymbol{p}_l {\boldsymbol{G}_l}^{-1} \boldsymbol{k}_l^\top - \sum_{m=1}^{M}\sum_{t=1}^{T} \gamma_{m,t} = 0 \quad (14)$$

where M is the total number of mixtures and $\gamma_{m,t}$ is the state occupancy probability of m-th mixture at time t. The $\boldsymbol{k}_l$ and $\boldsymbol{G}_l$ parameters are given by.

$$\boldsymbol{k}_l = \sum_{m=1}^{M} \frac{1}{\sigma_{m_l}^2} \mu_{m_l} \sum_{t=1}^{T} \boldsymbol{\gamma}_{m,t} \boldsymbol{\xi}_t^T + \tau \boldsymbol{h}_l \quad (15)$$

$$= \boldsymbol{k}_{\mathrm{ML}} + \tau \boldsymbol{h}_l \quad (16)$$

$$\boldsymbol{G}_l = \sum_{m=1}^{M} \frac{1}{\sigma_{m_l}^2} \sum_{t=1}^{T} \boldsymbol{\gamma}_{m,t} \boldsymbol{\xi}_t \boldsymbol{\xi}_t^T + \tau \boldsymbol{I}_L \quad (17)$$

$$= \boldsymbol{G}_{\mathrm{ML}} + \tau \boldsymbol{I}_L \quad (18)$$

where $\boldsymbol{h}_l$ is the l-th row of the matrix $\boldsymbol{H}$. μ_{m_l} and $\sigma_{m_l}^2$ are the l-th element of the mean vector of the m-th mixture and the diagonal element of covariance matrix of the m-th mixture, respectively. From these equations, we can see that hyperparameters $\boldsymbol{h}_l$ and $\boldsymbol{I}_L$ smooth the ML statistics $\boldsymbol{k}_{\mathrm{ML}}$ and $\boldsymbol{G}_{\mathrm{ML}}$.

C. Combining VTLN with CMLLR

There have been many attempts to combine VTLN with other linear transformations. One of the first was by Pye and Woodland [17] to combine VTLN with MLLR transforms for speaker adaptive training. VTLN was shown to give additive performance. It was mentioned by Uebel and Woodland [18] that estimating both transforms would be no better than just using CMLLR unless the effect of initialization is of key importance. The combination of VTLN and CMLLR can give

additional improvements only in special situations where the initialization of the transformation is important. When multiple iterations of CMLLR transform estimation is performed, the combination does not give any additional improvements. The same reason was postulated for having additional performance improvements after multiple iterations of CMLLR. It was also shown by Panchapagesan and Alwan [19] that estimating a bias vector and unconstrained variance transformation on top of the linear transform based frequency warping can further improve the recognition accuracy. This phenomenon is mainly observed with very limited adaptation data (of the order of one adaptation sentence) compared to the MLLR transforms which outperform with more adaptation data.

Breslin et al. [11] showed that VTLN can be combined with CMLLR for rapid adaptation in ASR. In that work, a count smoothing framework is used to incorporate the prior information. The count smoothing framework was initially presented by Flego and Gales [20], where the predictive and adaptive noise compensating transforms were combined using this scheme. The predictive approaches make use of a mismatch function that represents the impact of the background noise on the clean speech. The number of parameters associated with this mismatch function is usually small. This is in contrast to adaptive approaches to speaker and noise compensation where, normally, a large number of linear transforms of the model parameters are estimated. Flego and Gales [20] mention that CMLLR does not have a conjugate prior, instead count smoothing can be used to combine it with the predictive transforms. The pseudo counts associated with the predictive transform are combined with the actual observed counts and the transforms are estimated.

Breslin et al. [11] used this count smoothing framework to combine rapid adaptation techniques such as VTLN and predictive CMLLR (pCMLLR) with CMLLR transforms. Statistics $\boldsymbol{k}_l$ and $\boldsymbol{G}_l$ for estimating the final transform are based on the interpolation between ML statistics of adaptation data and prior statistics obtained from VTLN or pCMLLR and are given by

$$\boldsymbol{k}_l = \boldsymbol{k}_{\text{ML}} + \tau \frac{\boldsymbol{k}_{\text{pri}}}{\sum_m \gamma_m} \tag{19}$$

$$\boldsymbol{G}_l = \boldsymbol{G}_{\text{ML}} + \tau \frac{\boldsymbol{G}_{\text{pri}}}{\sum_m \gamma_m} \tag{20}$$

The prior statistics $\boldsymbol{k}_{\text{pri}}$ and $\boldsymbol{G}_{\text{pri}}$ are normalized so that they effectively contribute τ frames to the final statistics. γ_m represents the state occupancy probabilities for the output distributions. As more data becomes available, the CMLLR statistics $\boldsymbol{G}_{\text{ML}}$ and $\boldsymbol{k}_{\text{ML}}$ will dominate, but for small amounts of data the prior statistics are more important.

III. Combining VTLN with CSMAPLR

This section explains the proposed method for combining VTLN with CSMAPLR and also presents a few alternative approaches for the combination.

A. Proposed method: VTLN prior for CSMAPLR

CSMAPLR uses a global transform based on the ML criterion at the root node and hence there is no prior distribution at the root node. However, if the amount of adaptation data is very limited, even the estimation of the global transform may suffer from the lack of data. For such cases, the MAP criterion can be used at the root node as well.

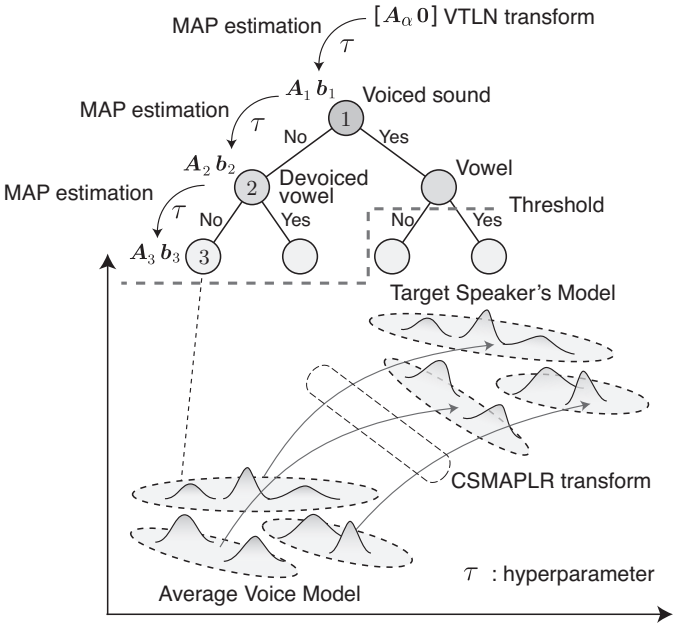


Fig. 1. CSMAPLR (VTLN prior): a global VTLN transform is used as a hyperparameter for the MAP estimation of the CMLLR transform at the root node of the regression class tree. The MAP-estimated CMLLR transform is then propagated to the child nodes as a hyperparameter for the MAP estimation at the child nodes in similar way to CSMAPLR.

A possible choice for the hyperparameter at the root node is the use of an identity matrix, that is, $\boldsymbol{H} = [\boldsymbol{I}_L, \boldsymbol{0}]$ and this can smooth the ML statistics at the root node. However, a better choice for the hyperparameter at the root node would be the use of a VTLN transform as suggested by Breslin et al. [11].

CSMAPLR uses the matrix variate normal distributions of Eq (12) as an approximated prior distribution. Although this is not a conjugate prior, this convenient prior allows us to directly use the VTLN transform as a hyperparameter at the root node, by setting the hyperparameter $\boldsymbol{H}$ representing the mean of the prior distribution as

$$\boldsymbol{H}_\alpha = [\boldsymbol{A}_\alpha, \boldsymbol{0}] \tag{21}$$

where $\boldsymbol{A}_\alpha$ is the VTLN transformation matrix described by α and $\boldsymbol{0}$ is a zero bias vector[2]. Figure 1 illustrates the proposed idea.

The VTLN transform may be used for the dynamic features of the cepstra; in this case the hyperparameter matrix $\boldsymbol{H}$ is a block diagonal matrix with repeating $\boldsymbol{A}_\alpha$ matrix.

$$\boldsymbol{H}_\alpha = \begin{bmatrix} \boldsymbol{A}_\alpha & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_\alpha & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_\alpha & \boldsymbol{0} \end{bmatrix} \tag{22}$$

[2]Instead of the zero bias vector, we may estimate the bias term $\boldsymbol{b}_0$ in addition to the the VTLN matrix and may set the hyperparameter $\boldsymbol{H}$ as $\boldsymbol{H}_\alpha = [\boldsymbol{A}_\alpha, \boldsymbol{b}_0]$. However adding the bias term to the hyperparameter $\boldsymbol{H}$ at the root node of the regression class tree did not show any noticeable improvements and hence we have decided to use the zero bias vector in the experiments described later.

TABLE I
COMPARISON AND DEFINITION OF TERMINOLOGIES USED FOR THE METHODS.

Criterion	Prior type at root node	Terminology
ML	Uniform	CMLLR
SMAP	Uniform	CSMAPLR
SMAP	Identity	CSMAPLR (Identity prior)
SMAP	VTLN	CSMAPLR (VTLN prior)

Compared to Eqs (16) and (19), the first-order statistics $\boldsymbol{k}_l$ are smoothed using the VTLN transform at the root node of the regression class tree as follows:

$$\boldsymbol{k}_l = \boldsymbol{k}_{\text{ML}} + \tau \boldsymbol{h}_{\alpha,l} \tag{23}$$

where $\boldsymbol{h}_{\alpha,l}$ is the l-th row of the $\boldsymbol{H}_\alpha$.

VTLN can capture the gender or age characteristics of a speaker. Hence, we expect that these characteristics captured by VTLN transform $\boldsymbol{H}_\alpha$ are better propagated to the nodes of the tree structure than the uniform distribution or the identity prior, and hence that it improves the speaker characteristics of adapted models even if the amount of adaptation data is very small. This proposed method is called "CSMAPLR (VTLN prior)". If the identity matrix is used as the initial prior for CSMAPLR at the root node instead of the VTLN transform, it is called "CSMAPLR (identity prior)." Please refer to Table I for the definition of terminologies of these methods.

There are pros and cons compared to the method proposed by Breslin et al. [11]. – In the proposed CSMAPLR (VTLN prior) approach, the VTLN transform is used only for smoothing of the first-order statistics $\boldsymbol{k}_l$ whereas the second-order statistics $\boldsymbol{G}_l$ are also smoothed in the approach by Breslin et al. [11], as shown in Eq (20). On the other hand, the proposed CSMAPLR (VTLN prior) approach uses the VTLN transform directly for smoothing of the statistics and hence there is no need to compute and store the prior statistics $\boldsymbol{k}_{\text{pri}}$. There should not be any performance difference for the proposed method when compared to the approach by Breslin et al. [11]. The advantage of the proposed method is that this requires less time and space complexity because VTLN transforms are directly used as priors. Moreover, method by Breslin et al. [11] uses a heuristic approach and this work presents a structured mathematical framework and derivation for combining the model and feature transformation. Since we do not expect any significant performance difference, and also knowing the fact that the proposed method has better time and space complexity, there is no comparison presented in this work between the two methods.

B. Other methods for combining VTLN with CSMAPLR

Here we describe other two methods for combining VTLN with CSMAPLR, which may be compared with the proposed CSMAPLR (VTLN prior).

It is possible to apply the VTLN transforms onto the average voice model first and then to apply CSMAPLR transforms further on the top of that. This is called "cascade" transform. It was shown by Karhila et al. [21] that VTLN and CSMAPLR cascade transformations can improve child synthetic speech where the adult average voice was transformed into child voice.

In the proposed CSMAPLR (VTLN prior) adaptation, the global VTLN transform is used explicitly only at the root node of the regression class tree and it is propagated into child node based on the SMAP criterion. Instead of the SMAP propagation, it is also possible to explicitly use VTLN transforms at each of the regression classes for the MAP estimation of the CMLLR transform. For simplicity this paper does not show the results of this approach. However, readers interested in this approach may refer [22].

IV. EVALUATIONS IN MATCHED CONDITIONS

This section shows experimental results of HMM-based TTS and ASR systems using the proposed technique in matched conditions.

A. HMM-based TTS

The HMM speech synthesis system (HTS) [1] was used for generating acoustic parameters for speech synthesis. HTS models spectrum, $\log F_0$, band-limited aperiodic components and duration in the unified framework of hidden semi-Markov models (HSMMs). The STRAIGHT vocoder [23] was used to synthesize speech waveforms from the acoustic parameters generated from the HSMMs. The HMM topology used was five-state and left-to-right with no skip states. Speech features were 59th-order mel-cepstra, $\log F_0$, 25-dimensional band aperiodicity, and their delta and delta-delta coefficients, extracted from 48kHz recordings with a frame shift of 5ms. The speaker-dependent model was built using a UK English speech corpus including 5 hours of clean speech data uttered by a RP[3] male professional narrator (source speaker). The first evaluation experiments were performed by adapting the speaker-dependent model to a different UK English male semi-professional speaker (target speaker) who has the same RP accent as the source speaker.

Objective evaluation based on the mel-cepstral distance (MCD) was carried out. The MCD is the Euclidean distance between the synthesized cepstra and those derived from the natural speech, and can be viewed as an approximation to the log spectral distortion measure according to Parserval's theorem. One hundred sentences were synthesized for measuring the average MCD.

In addition, the subjective listening tests were performed by 17 subjects using the Blizzard challenge 2010 test sentences for naturalness, speaker similarity and intelligibility with different amounts of adaptation data. The techniques compared in this experiment were VTLN, CSMAPLR and CSMAPLR (VTLN prior) systems.

The subjective tests were based on mean opinion scores (MOS) for the naturalness, ABX scores for the speaker similarity, word error rate (WER) for the intelligibility. The synthesized utterances were rated on a 5-point scale for the MOS test, 5 being "completely natural" and 1 being "completely

[3]According to the Oxford English dictionary Received Pronunciation (RP) refers to a standard accent of English as spoken in the South of England

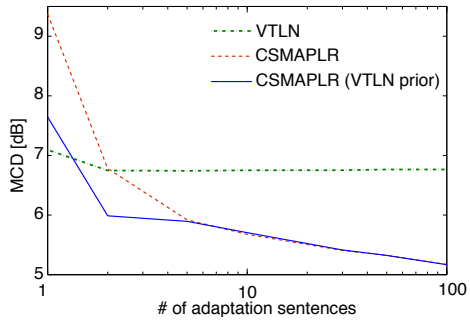


Fig. 2. Mel-cepstral distances between reference speech and synthetic speech adapted using VTLN, CSMAPLR, and CSMAPLR (VTLN prior) in the matched condition.

unnatural". The source and the target speakers were given as the two reference speakers in the ABX test and the subjects were asked to compare speaker similarity of synthetic speech with these references. For the intelligibility test, semantically unpredictable sentences were used and the subjects were asked to type what they have heard. In these listening tests, only the spectral parameters were adapted and other excitation and duration parameters were not adapted so that the subjects can pay attention to the spectral differences.

1) Objective Evaluation: The values of the MCD for different amounts of adaptation data are plotted in the Figure 2. The objective results show that 1) the VTLN technique works best in comparison to others when one adaptation sentence is used (around 7dB), whereas its performance does not improve if more than one sentence is used; and that 2) the CSMAPLR improves the MCD to around 5dB when the number of adaptation sentences is more than five. However, the performance of the CSMAPLR technique rapidly becomes worse when the number of adaptation sentences is less than five, reaching around 9.5dB MCD with only one adaptation utterance. Finally, the objective results clearly show that the proposed CSMAPLR (VTLN prior) technique alleviates this issue of the CSMAPLR technique and improves the performance when the number of adaptation sentences is less than five. We can see that even if the number of adaptation sentences is just two, the performance of the CSMAPLR (VTLN prior) technique outperforms the VTLN technique; its distortion is around 6dB.

2) Subjective Evaluation: In the subjective listening tests, synthetic speech utterances generated from the models adapted using 1, 10 and 100 sentences were compared by the subjects. The results of the listening tests are shown in Figure 3, which is, from left to right, the mean opinion scores for the naturalness, the ABX scores for the similarity to the target speaker, and WER for the intelligibility.

a) Mean opinion score (Naturalness): From the mean opinion scores (MOS) on naturalness, we first see that CSMAPLR has the worst MOS value 1.1 when the number of adaptation sentences is one. The MOS value of the CSMAPLR approach become better as the number of adaptation sentences increases. We then see that the MOS of the VTLN approach are high (3.4), however, VTLN does not improve naturalness significantly even if more data is used. Finally we can see that VTLN prior is useful for CSMAPLR. Using the VTLN transform as an initial prior for CSMAPLR, the MOS value increases from 1.1 to 1.9 when the number of adaptation sentences is one. When the number of adaptation sentences is ten, the VTLN prior increases the MOS value from 3.0 to 3.2. There was no difference when 100 sentences were used.

b) ABX score (Similarity): From the ABX scores, which are percentages of synthetic speech utterances that were judged by the subjects as closer to the target speaker compared to the source speaker, we can first see that when the number of adaptation sentence is one, synthetic speech using CSMAPLR was judged to be similar to neither source nor target speakers. When the number of adaptation sentences is ten or hundred, synthetic speech using CSMAPLR was judged to be similar to target speakers. This is consistent with the results of naturalness evaluation above. We then see that when the number of adaptation sentences is one, synthetic speech using VTLN was judged to be similar to the target speaker. However, the ABX scores of the VTLN approach do not increase even if more data is used. Finally we see that the CSMAPLR (VTLN prior) has a better ABX score than the CSMAPLR without the VTLN prior when the number of adaptation sentence is one. However the VTLN prior did not improve the ABX scores when the number of adaptation sentences is ten or hundred. This is probably because these experiments are in matched conditions and the source and target speakers are similar to one another to some extend. This will not be the case in a mis-matched condition where the source and target speakers are very different like a child speech evaluated on a male speaker model. Such a scenario can illustrate the limitation of VTLN in capturing speaker characteristics with a single transformation parameter. These experiments are presented in section V-A with age transform based mis-matched train and test conditions.

c) WER (Intelligibility): From the intelligibility evaluation, we first observe that CSMAPLR has significantly degraded intelligibility with one adaptation sentence and that the proposed CSMAPLR (VTLN prior) technique alleviates this issue. The proposed method is able to preserve the intelligibility of a VTLN adapted system.

B. HMM-based ASR

Following the work by Dines et al. [24, 25], this section presents ASR experiments to show that the proposed techniques can be used for both TTS and ASR equally well. It should be noted that the results may not be the state-of-the-art for this corpus since our strategy is to use similar models for ASR and TTS in accordance with the unification theme in a speech-to-speech translation system.

The hidden Markov models were built with 13 dimensional cepstral features with Δ and Δ^2 for the (US English) WSJ0 database. The spectral features were extracted using STRAIGHT. Speech recognition and synthesis systems use the same average voice training procedure, which involves speaker adaptive training (SAT) and context clustering using decision trees. The experimental set-up is the same as that of Dines

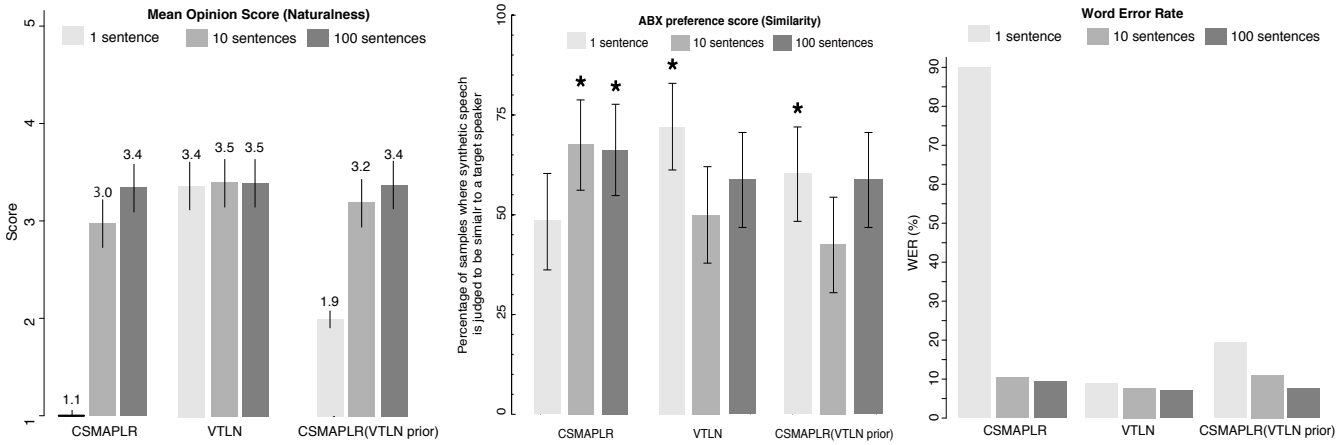


Fig. 3. Listening tests results. There are three columns of plots which are, from left to right, mean opinion score for naturalness, similarity to target speaker, and intelligibility. Bars of mean preference scores and ABX preference scores indicate 95% confidence intervals based on t-distribution. * of the ABX preferences scores for speaker similarity indicates a system where synthetic speech generated from adapted models was judged as closer to the target speaker compared to the source speaker ($p < 0.05$).

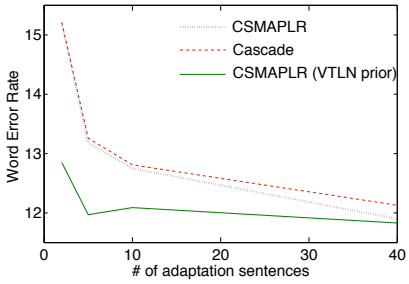


Fig. 4. Word error rate for VTLN in combination with CSMAPLR systems for (s4-c3) Nov93 evaluations of WSJ0.

et al. [24][4].

The WERs using different amounts of adaptation data ranging from 2 to 40 adaptation sentences are shown in Figure 4. In the figure, cascade transforms of VTLN and CSMAPLR are shown in addition to CSMAPLR with and without the VTLN prior. It can be observed that the VTLN prior provides marked improvements to the ASR performance of CSMAPLR, especially when the adaptation data is limited, but, it still performs slightly better even for 20 to 30 adaptation sentences. We did not observe significant gains with the cascade transform of VTLN and CSMAPLR transforms from this experiment.

V. Evaluations in Mismatched Conditions

This section shows additional experimental results of HMM-based TTS and ASR systems using the proposed technique in special conditions where target speakers are not matched with training speaker in terms of age, gender and recording environments. In a mismatched condition, the CSMAPLR transform tends to capture the mismatch as well and not just speaker specific characteristics. It can be postulated that using VTLN along with CSMAPLR will restrict the CSMAPLR transforms to capture only the speaker specific characteristics and will yield better performance especially when the amount of adaptation data is limited.

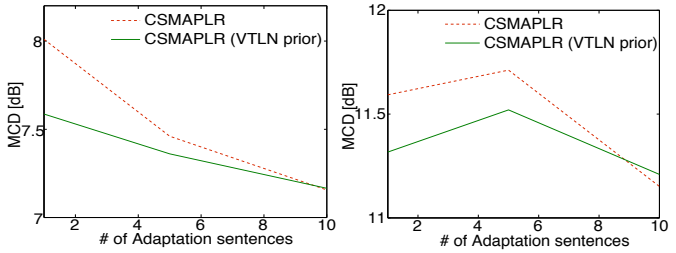
A. Age Transforms

The vocal tract length is proportional to the actual size of the individual and hence, is shortest in a child. The details of the vocal tract length being proportional to the actual body size and the differences in vocal tract length in growing children of different age-groups was investigated by Hancil and Hirst [26] and also by Fitcha and Giedd [8].

This poses a case of extreme frequency warping when an adult, particularly, male model is adapted to a child voice. The VTLN prior representing the vocal tract length should give performance improvements for such a situation. The influence of VTLN was evaluated using both subjective and objective evaluations for speech synthesis. The speech sampled at 48kHz were collected in-house from two different children at the anechoic recording studio of the Centre for Speech Technology Research (CSTR), Edinburgh. The children were asked to read fairy tales. Only the data from one child was manually annotated to have a full set of reference data for objective evaluations. Other child had only four annotated sentences for adaptation. This child was used in the subjective evaluations to clarify the effects seen with the earlier subjective evaluations in section IV-A2 where the test and train speakers sound very similar. In this case, the adult model is transformed to child speech and results for speaker similarity should clarify the limitations of VTLN in capturing speaker characteristics.

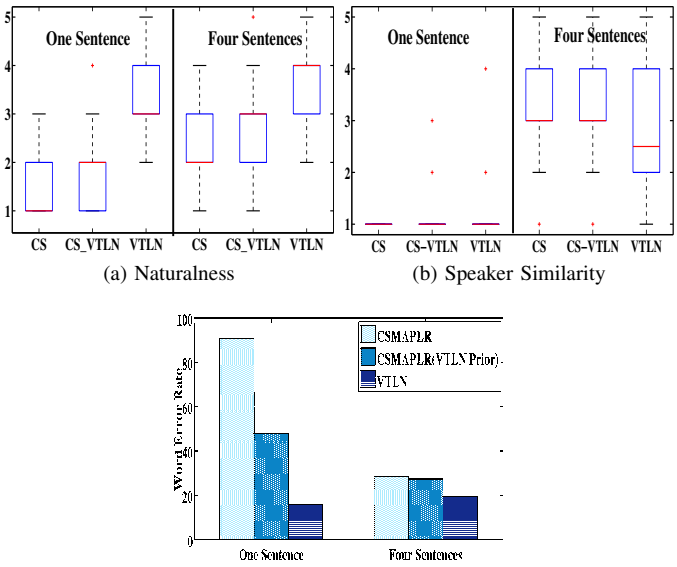
For the objective evaluations, speech utterances taken from the corpus were used to adapt gender dependent average voice models, which were trained using speech data uttered by about 17 adult male or 19 adult female speakers, respectively. Experiments were carried out with different amounts of adaptation data in order to adapt the male and female average voice models to the child voice. In a similar way to the previous experiment, we have generated a hundred synthetic speech utterances and have measured the MCD of them in an objective measure.

[4]The baseline system is the system 'd' in Table IX of [24], which has 13% word error rate (WER). The baseline system reported in [24] uses the value of τ, the weight of the prior as one. Increasing this value to 1000 improves the WER of CSMAPLR up to 12%.



(a) Female Average Voice Model (b) Male Average Voice Model

Fig. 5. MCD between child synthetic speech and reference natural speech. Both male and female adult average voice models were adapted to a child voice.



(a) Naturalness (b) Speaker Similarity

(c) Intelligibility

Fig. 6. Subjective evaluations for child speech adaptation on a adult male speaker model. CS refers to CSMAPLR and CS-VTLN refers to CSMAPLR (VTLN Prior).

Figure 5 shows the MCD between child synthetic speech and reference natural speech. As expected, the MCD of child voice adapted from the male average voice model has significantly larger distortion than that adapted from the female average voice model because of the difference of vocal tract length. Then we can see that the proposed VTLN prior has very good influence when the amount of adaptation data is as little as one sentence. MCD has reduced to about 7.5 dB from 8 dB for the case of the female average voice model. As the amount of adaptation data increases, the difference between CSMAPLR with and without the VTLN prior becomes smaller. It can be observed that there is a slight incidental degradation of performance with 10 sentence adaptation on Male Average Voice model. The difference is less than 0.1dB and not perceivable. Also, these are mean MCD values across the test utterances, the variance of MCD scores is higher for CSMAPLR (1.8824 for CSMAPLR and 1.8247 for CSMAPLR (VTLN Prior)) in this case which further rules out any significant difference.

The speaker dependent male model (same as in Section IV-A) is used as the base model for adapting to the child speech. The subjective evaluations for naturalness, speaker similarity and intelligibility were performed on three different systems: VTLN, CSMAPLR and CSMAPLR (VTLN prior). All systems are evaluated for adaptation with one and four sentences. 14 listeners participated in these evaluations and the results are plotted in Figure 6. The results for naturalness and speaker similarity are plotted as MOS ranging from 1 (Completely Unnatural / Sounds like a totally different person) to 5 (Completely Natural / Sounds like exactly same person). The word error rates for the text typed in by the listeners after perceiving the target speech is plotted as the result for speech intelligibility. It can be observed from the results that VTLN gives the best naturalness and intelligibility scores especially for a single sentence adaptation. In this case, CSMAPLR transformation is not intelligible at all. This gives further proof to the hypothesis that VTLN is useful as a rapid adaptation performance in child speech synthesis. The proposed method, CSMAPLR(VTLN prior), has better speaker similarity compared to VTLN with four adaptation sentences as opposed to the observations in earlier subjective evaluations in the matched conditions. The speaker similarity given by four sentence VTLN adaptation is due to the naturalness of the synthesised speech. As observed in our previous studies, listeners were judging naturalness instead of speaker similarity. This can be validated by listening to the samples in the demonstration page: www.idiap.ch/~lsaheer/VTLNSMAP/demo.html.

B. Cross-Gender Transforms

Another mismatched scenario where VTLN may perform better is the wider variation of the vocal tract length of speakers used for training and adaptation. This is more critical when we try to adapt gender-dependent average voice models to other genders. There may be only subtle changes in the vocal tract length within the same gender, especially, the differences less than the value of 0.2 for the warping factor may not be perceivable. Across genders where the difference in vocal tract length is significant, this factor alone may be able to represent the target speaker to some extent even if there are more speaker specific pitch and other characteristics ignored.

For this purpose, we have used the CSTR VCTK corpus[5]. This corpus was recorded at the Centre for Speech Technology Research (CSTR), University of Edinburgh, UK in a specialized anechoic recording room and has speech data uttered by 109 native speakers of English with various accents. From this corpus, we have chosen 31 male and 29 female native speakers of UK English as target speakers and have adapted the UK English gender-dependent average voice models to them to see the impact of the VTLN prior from many speakers, especially in cross-gender cases. The gender-dependent average voice models were the same as those used for the age transforms previously mentioned.

A randomly chosen single adaptation sentence was used to generate the transforms for each method. In a similar way to previous experiments, 100 sentences were synthesized with each of these techniques for each of test speakers and the MCD was measured from the synthetic speech utterances as the

[5]http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html

TABLE II
CROSS-GENDER SPEAKER ADAPTATION EXPERIMENTS FOR SPEECH SYNTHESIS. THE AVERAGE MCD WAS CALCULATED USING ABOUT 30 SPEAKERS FOR EACH GENDER USING A SINGLE SENTENCE AS ADAPTATION DATA FOR EACH SPEAKER.

Method	Male AVM to males	Male AVM to females	Female AVM to males	Female AVM to females
CMLLR	6.6	8.8	7.1	6.6
CSMAPLR	6.4 (-0.2dB)	8.6 (-0.2dB)	6.7 (-0.4dB)	6.4 (-0.2dB)
CSMAPLR (Identity prior)	6.4 (-0.2dB)	8.4 (-0.4dB)	6.8 (-0.3dB)	6.4 (-0.2dB)
CSMAPLR (VTLN prior)	6.4 (-0.2dB)	8.3 (-0.5dB)	6.6 (-0.5dB)	6.3 (-0.3dB)
Cascade	6.7 (+0.1dB)	9.2 (+0.4dB)	7.5 (+0.4dB)	6.7 (+0.1dB)

objective measure. For the full comparison, we have computed the MCD of 1) CMLLR, 2) CSMAPLR, 3) CSMAPLR with the identity prior at the root node, 4) CSMAPLR with the VTLN prior at the root node, and 5) the cascade transforms of VTLN and CSMAPLR.

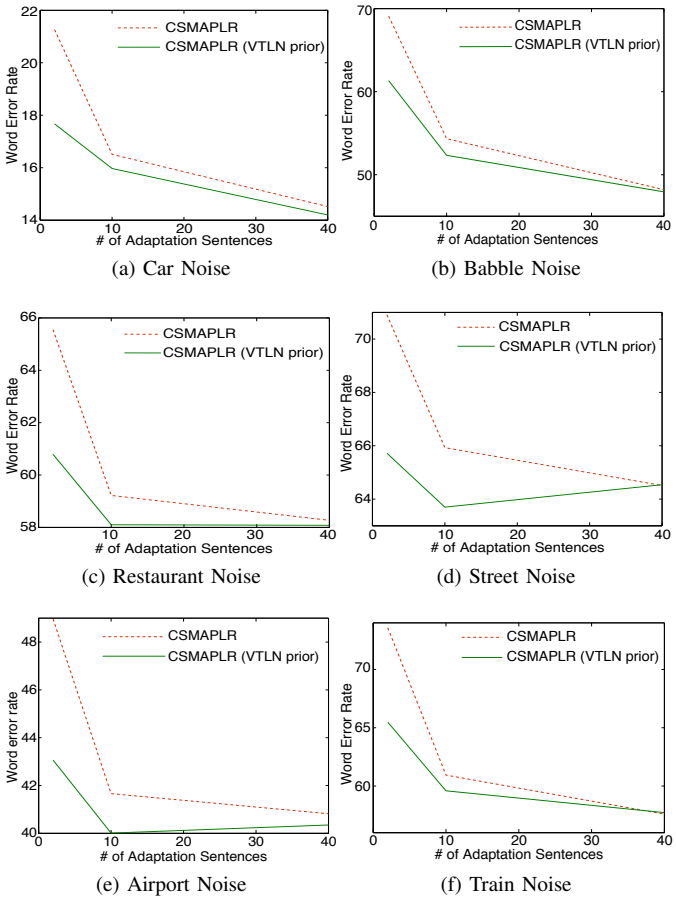
The objective results are shown in Table II. The table shows the performance of the male and female test speakers with each of the gender dependent average voice models (AVM). The results show that the VTLN prior for CSMAPLR gives the lowest MCD value and overall best performance in all cases. Interestingly, the identity prior for CSMAPLR also performs well in all cases. This emphasizes the fact that the initial prior is an important factor for the CSMAPLR transforms and further, an appropriate choice of the prior such as the VTLN prior can further improve performance. VTLN accounts for gender characteristics and is important in the case of cross-gender transformations. A VTLN prior in this case can provide more information and results in more significant reductions of the MCD scores compared to the same gender case. For example the VTLN prior resulted in 0.5dB reduction compared to CMLLR and 0.3dB reduction compared to CSMAPLR for the male AVM to female adaptation case, which are larger reductions than those for the male AVM to male adaptation. This is also true for the female AVM to male adaptation compared to the female AVM to female adaptation.

C. Noise-robust ASR and TTS

The final special scenario where VTLN is hypothesised to be important is the use of adaptation data that has varying background noise. Since both CMLLR and CSMAPLR do not have strong constraints in the affine transforms, this may overfit the models to the background noise of the adaptation data. It may be a good strategy to impose the VTLN prior in order to avoid such overfitting and thus, handle the varying background noise of the test data better. Both noise-robust ASR and TTS experiments are presented in this section to validate this hypothesis.

1) Noise-Robust ASR: The Aurora4 database represents a noisy speech data version of the WSJ0 database. The ASR models built using the WSJ database were used to recognize noisy speech data taken from the Aurora database. Similar to the ASR experiments presented in Section IV, the models were trained in accordance with the unification theme for ASR and TTS – The 13 dimensional MGCEP coefficients were used to generate the HMMs. Again, the experimental setup was same as that in [24].

There are six different noise types in the Aurora4 database: car noise, babble noise, street noise, airport noise, restaurant noise, and train noise. Evaluations were performed using different amounts of adaptation data ranging from 2 to 40 adaptation sentences in each of the noise conditions and we have compared the CSMAPLR with and without the VTLN prior.



(a) Car Noise (b) Babble Noise (c) Restaurant Noise (d) Street Noise (e) Airport Noise (f) Train Noise

Fig. 7. WERs calculated on the Aurora4 database including 6 types of noises. The ASR models were trained on the WSJ database.

The results are plotted in Figure 7. It can be seen from the results that the VTLN prior for CSMAPLR gives considerable improvements in the presence of all types of noises, especially when the amount of adaptation data is less than ten sentences. As more adaptation data comes in, the prior does not have much effect.

As mentioned earlier, it is worth noting that the overall performance presented here cannot be compared with the state-of-the-art results. This is due to the fact models are aligned to the unification of TTS and ASR and not exactly the perfect setup for an ASR system. Furthermore, no noise reduction/compensation techniques (like the methods in [27])

TABLE III
NOISE-ROBUST SPEAKER ADAPTATION EXPERIMENTS FOR SPEECH SYNTHESIS USING SINGLE SENTENCE AS ADAPTATION DATA FOR EACH SPEAKER. THE AVERAGE MCD WAS CALCULATED USING 45 SPEAKERS WHO WERE RECORDED AT A PUBLIC SPACE IN THE PRESENCE OF BABBLE NOISE.

Method	Male AVM	Female AVM
CMLLR	7.0	7.2
CSMAPLR	7.0 (0.0dB)	7.1 (-0.1dB)
CSMAPLR (Identity prior)	6.7 (-0.3dB)	6.9 (-0.3dB)
CSMAPLR (VTLN prior)	6.7 (-0.3dB)	6.8 (-0.4dB)
Cascade	11.3 (+4.3dB)	7.5 (+0.2dB)

were applied for improving the performance in noisy conditions. In this paper we only aim for proving that the techniques developed for TTS can be plugged into ASR in order to improve the performance.

2) Noise-robust TTS: As a part of the EC FP7 EMIME project, speech synthesis data for noise-robust TTS systems was collected at a conference venue (Interspeech 2009). Participants were asked to read aloud some texts in the background of the conference hall. This resulted in speech synthesis data uttered by 39 male and 6 female speakers in the presence of strong and varying babble noise. Using the noisy speech data, we have adapted the gender-dependent average voice models trained on clean speech data that were also used in the age transform and cross-gender experiments in previous sections. In a similar way to the cross-gender experiments, we have adapted the average voice models using a single sentence as adaptation data, have synthesized the 100 synthetic speech utterances for each of the five methods (and for each of test speakers), and have computed the MCD from the utterances as the objective measure.

The objective results are shown in Table III. The results are consistent with the observations made earlier with cross-gender experiments and show that the identity prior for CSMAPLR performs well and that the VTLN prior for CSMAPLR gives the lowest MCD value overall.

VI. CONCLUSIONS

This paper has presented a novel approach to combine the merits of VTLN and CSMAPLR, resulting in an improved adaptation technique for both HMM-based ASR and TTS. The proposed method is an efficient algorithm to smooth the first-order statistics required for the CSMAPLR using the VTLN transform directly. We conclude that the proposed VTLN prior for CSMAPLR can significantly improve the adaptation performance when the adaptation data is very limited. In HMM-based TTS, the proposed method improved naturalness and intelligibility of HMM-based synthetic speech compared to that using the CSMAPLR without the VTLN prior. In HMM-based ASR the proposed method performs better than the cascade transforms of VTLN and CSMAPLR. Performance improvements were also confirmed for mismatched conditions, especially when very little adaptation data was available. It was showed that the VTLN prior had led to the improvements of CSMAPLR adaptation when the test data has the mismatched conditions in terms of age, gender, and recording environments.

REFERENCES

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.
[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] L. Saheer, J. Dines, and P. N. Garner, "Vocal tract length normalization for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 7, pp. 2134–2148, 2012.
[4] T. Toda and K. Tokuda, "Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory HMM," in *Proc. of ICASSP*, Las Vegas, Nevada, USA, 2008, pp. 3925–3928.
[5] X. Zhuang, Y. Qian, F. K. Soong, Y.-J. Wu, and B. Zhang, "Formant-based frequency warping for improving speaker adaptation in HMM TTS," in *Proc. of Interspeech*, Makuhari, Japan, 2010, pp. 817–820.
[6] J. W. McDonough, "Speaker compensation with all-pass transforms," Ph.D. dissertation, John Hopkins University, 2000.
[7] D. Y. Kim, S. Umesh, M. J. F. Gales, T. Hain, and P. C. Woodland, "Using VTLN for broadcast news transcription," in *Proc. of ICSLP*, Jeju Island, Korea, 2004, pp. 1953–1956.
[8] W. T. Fitcha and J. Giedd, ""morphology and development of the human vocal tract: A study using magnetic resonance imaging," *Journal of Acoustical Society of America*, vol. 106, no. 3, September 1999.
[9] 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.
[10] L. Saheer, P. N. Garner, J. Dines, and H. Liang, "VTLN adaptation for statistical speech synthesis," in *Proc. of ICASSP*, Dallas, Texas, USA, Mar. 2010, pp. 4838–4841.
[11] C. Breslin, K. Chin, M. Gales, K. Knill, and H. Xu, "Prior information for rapid speaker adaptation," in *Proc. of Interspeech*, Makuhari, Japan, 2010, pp. 1644–1647.
[12] O. Shiohan, T. Myrvoll, and C. Lee, "Structural maximum a posteriori linear regression for fast HMM adaptation," *Computer, Speech and Language*, vol. 16, no. 3, pp. 5–24, Jan. 2002.
[13] W. Chou, "Maximum a posterior linear regression with elliptically symmetric matrix variate priors," in *Proc. of Eur. Conf. Speech Communication Technology*, Budapest, Hungary, 1999.
[14] 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, Yokohama, Japan, Sep. 1994, pp. 1043–1046.
[15] 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.
[16] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech & Language*, vol. 12 (2), pp. 75–98, 1998.
[17] D. Pye and P. C. Woodland, "Experiments in speaker normalisation and adaptation for large vocabulary speech recognition," in

Proc. of ICASSP, Munich, Bavaria, Germany, 1997, pp. 1047–1050.

[18] 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, Eurospeech*, Budapest, Hungary, 1999, pp. 2527–2530.

[19] 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.

[20] F. Flego and M. J. F. Gales, "Incremental predictive and adaptive noise compensation," in *Proc. of ICASSP*, Washington, DC, USA, 2009, pp. 3837–3840.

[21] R. Karhila, R. S. Doddipatla, M. Kurimo, and P. Smit, "Creating synthetic voices for children by adapting adult average voice using stacked transformations and VTLN," in *Proc. of ICASSP*, Kyoto, Japan, March 2012, pp. 4501–4504.

[22] L. Saheer, "A unified framework of feature based adaptation for statistical speech synthesis and recognition," Ph.D. dissertation, Idiap Research Institute, Ecole Polytechnique de Lausanne, 2013.

[23] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[24] J. Dines, J. Yamagishi, and S. King, "Measuring the gap between HMM-based ASR and TTS," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1046–1058, Dec. 2010.

[25] J. Dines, L. Saheer, and H. Liang, "Speech recognition with synthesis models by marginalising over decision tree leaves," in *Proc. of Interspeech*, Brighton, UK, Sep. 2009, pp. 1395–1398.

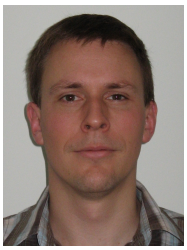
[26] S. Hancil and D. Hirst, *Prosody and Iconicity*. John Benjamins Publishing, 2013.

[27] P. N. Garner, "Cepstral normalisation and the signal to noise ratio spectrum in automatic speech recognition," *Speech Communication*, vol. 53, no. 8, pp. 991–1001, Oct. 2011.

Lakshmi Saheer recieved B.Tech degree in Computer Science from the Cochin University of Science & Technology, India in 2002 and the M.S degree in Computer Science from the Indian Institute of Technology Madras, India in 2006. She worked with Siemens Information Systems Ltd. (India), V-Enable Technologies (U.S.A.) and Sony Ericsson Mobile Communications (Sweden). She completed her doctoral studies combinely at the Idiap Research Institute, Martigny and the Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland in 2012. She is currently pursuing a post-doc at Idiap research institute. Her research interests include speech recognition, statistical parametric speech synthesis, feature normalization and speaker adaptation. She is a member of IEEE and a reviewer for IEEE Transactions on Audio, Speech and Language Processing and the Eurasip Journal.

Junichi Yamagishi is a lecturer and holds an EPSRC Career Acceleration Fellowship in the Centre for Speech Technology Research (CSTR) at the University of Edinburgh. He is also an associate professor of National Institute of Informatics (NII) in Japan. He was awarded a Ph.D. by Tokyo Institute of Technology in 2006 for a thesis that pioneered speaker-adaptive speech synthesis and was awarded the Tejima Prize as the best Ph.D. thesis of Tokyo Institute of Technology in 2007. Since 2006, he has been in CSTR and has authored and co-authored about 100 refereed papers in international journals and conferences. His work has led directly to three large-scale EC FP7 projects and two collaborations based around clinical applications of this technology. A recent coauthored paper was awarded the 2010 IEEE Signal Processing Society Best Student Paper Award and cited as a "landmark achievement of speech synthesis." He was awarded the Itakura Prize (Innovative Young Researchers Prize) from the Acoustic Society of Japan for his achievements in adaptive speech synthesis. In 2012 he was an area chair for the Interspeech conference and elected to membership of the IEEE Signal Processing Society Speech & Language Technical Committee. He is an external member of the Euan MacDonald Centre for Motor Neurone Disease Research. He has been Principal Investigator at Edinburgh on EPSRC and JST projects totalling over £1.8m.

Philip N. Garner Philip N. Garner received the degree of M.Eng. in Electronic Engineering from the University of Southampton, U.K., in 1991, and the degree of Ph.D. (by publication) from the University of East Anglia, U.K., in 2012. He first joined the Royal Signals and Radar Establishment in Malvern, Worcestershire working on pattern recognition and later speech processing. In 1998 he moved to Canon Research Centre Europe in Guildford, Surrey, where he designed speech recognition metadata for retrieval. In 2001, he was seconded (and subsequently transfered) to the speech group at Canon Inc. in Tokyo, Japan, to work on multilingual aspects of speech recognition and noise robustness. As of April 2007, he is a senior research scientist at Idiap Research Institute, Martigny, Switzerland, where he continues to work in research and development of speech recognition, synthesis and signal processing. He is a senior member of the IEEE, and has published internationally in conference proceedings, patent, journal and book form as well as serving as coordinating editor of ISO/IEC 15938-4 (MPEG-7 Audio).



John Dines graduated with first class honours in Electrical and Electronic Engineering from University of Southern Queensland in 1998 and received the Ph.D. degree from the Queensland University of Technology in 2003 with the thesis: "Model based trainable speech synthesis and its applications". Since 2003 he has been employed at the Idiap Research Institute, Switzerland, where he has been working mostly in the domain of meeting room speech recognition. A major focus of his current research is combining his background in speech recognition and speech synthesis to further advance technologies in both domains. He is a member of IEEE and a reviewer for IEEE Signal Processing Letters and IEEE Transactions on Audio, Speech and Language Processing.