COMBINING VOCAL TRACT LENGTH NORMALIZATION WITH HIERARCHIAL LINEAR TRANSFORMATIONS

Lakshmi Saheer [1,2], *Junichi Yamagishi*[3], *Philip N. Garner* [1], *John Dines* [1]

[1] Idiap Research Institute, Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[3] Centre for Speech Technology Research, University of Edinburgh, U.K.

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 proposes that the merits of VTLN can be combined with those of linear transform based adaptation in a hierarchial Bayesian framework, where VTLN is used as the prior information. A novel technique for propagating the gender information from the VTLN prior through constrained structural maximum a posteriori linear regression (CSMAPLR) adaptation is presented. Experiments show that the resulting transformation has improved speech quality with better naturalness, intelligibility and improved speaker similarity.

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

1. 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 previously developed for automatic speech recognition (ASR) [2]. 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 a considerable amount of adaptation data (of the order of 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]. By contrast, this technique requires very little adaptation data as it estimates only a single parameter. This system 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 [4–6]. 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. The formant frequency positions are inversely proportional to VTL, and hence can vary around 25% [7]. 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. [3, 8].

Breslin et.al. [9] showed that VTLN can be combined with constrained MLLR (CMLLR) for rapid adaptation in ASR. In that work, a count smoothing framework is used to incorporate the prior information. Structural maximum a posteriori (SMAP) based adaptation techniques also use prior information for transform estimation [10]. The SMAP technique uses a family of elliptically symmetric distributions including the matrix variate normal prior density as a prior distribution [11] 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 the context of statistical parametric speech synthesis.

Although CSMAPLR uses the identity matrix as a hyper parameter of the prior distribution at the root node, in a similar spirit to the work of Breslin et. al. [9], the hyper parameter at the root node may be replaced by a VTLN transform. The structural framework helps propagate the prior information affected by the VTLN transform through the various levels of the regression tree effectively. The tree structure is generated using linguistic information; hence, the propagated prior information should reflect the connection and similarity of the distributions of linguistic information. Using the VTLN matrix as the initial prior information for the root node of the CSMAPLR transform could result in the propagation of speaker characteristics and improved speaker adaptation even when very lit-

Thanks to FP7 Emime project for funding.

tle data is available.

2. THEORY

2.1. VTLN in Statistical Parametric Speech Synthesis

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) [12], 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 [13] 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:

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

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 [13]:

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

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

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

We may represent the linear 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. 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 [7]:

$$\widehat{\alpha}_s = \operatorname*{argmax}_{\alpha} P(\boldsymbol{x}_{\alpha_s} \mid \Theta, \alpha_s, w_s) \tag{9}$$

where $\boldsymbol{x}_{\alpha_s}$ represents features warped with the warping factor α_s for speaker s; 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 CMLLR transform 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 [3, 5, 14], which transforms the spectral features as follows

$$\tilde{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} = \boldsymbol{W}\boldsymbol{\xi}. \tag{10}$$

where $\boldsymbol{\xi} = [\boldsymbol{x}^\top, 1]^\top$, 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.

2.2. CSMAPLR

Constrained structural MAP based linear regression (CSMAPLR) is a robust framework to estimate the CMLLR transforms $\boldsymbol{W}$ based on the SMAP criterion [15]:

$$\widehat{\boldsymbol{W}}_s = \operatorname*{argmax}_{\boldsymbol{W}} P(\boldsymbol{x} \mid \Theta, \boldsymbol{W}_s, w_s)\, P(\boldsymbol{W}_s) \tag{11}$$

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

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

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 SMAP criterion, the tree structures of the distributions effectively control these hyperparameters. The whole adaptation data is used to estimate a global transform at the root node of the tree based on the ML criterion and it is propagated to the child nodes as a hyperparameter $\boldsymbol{H}$. The transforms at each child node are estimated using the corresponding adaptation data and hyperparameters propagated with the MAP criterion. This process is continued recursively from the root node to all the leaf nodes of the tree structure.

In the CSMAPLR estimation, the hyperparameter $\boldsymbol{\Psi}$ is fixed to the identity matrix and $\boldsymbol{\Omega}$ to a scaled identity matrix, $\boldsymbol{\Omega} = \tau_b \boldsymbol{I}_L$. τ_b is a positive scalar that controls the scale factor for the prior propagation and $\boldsymbol{I}_L$ is $L \times L$. The hyperparameter of the prior distribution $\boldsymbol{H}$ at the root node of the tree structure is set to an identity matrix, that is, a prior favouring no occupancy smoothing.

2.3. Using VTLN as CSMAPLR Prior

The VTLN transformation presented in this paper can be considered as a very constrained form of CMLLR/CSMAPLR. The single parameter normally gives some measure of the vocal tract length,

[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.

but more concretely is known to be highly correlated with basic speaker characteristics such as gender and as such can act as a prior for speaker independent modelling. In fact the CSMAPLR adaptation technique can use any arbitrary prior information (instead of the identity matrix) at the root node of the tree structure. This prior information can easily be replaced with the VTLN transformation matrix. At the root node, we may set the hyperparameter $\boldsymbol{H}$ as

$$\boldsymbol{H}_{\text{VTLN}} = [\boldsymbol{A}_\alpha, \boldsymbol{0}] \tag{13}$$

where $\boldsymbol{A}_\alpha$ is the VTLN transformation matrix described by α and $\boldsymbol{0}$ is a zero bias vector. The VTLN prior 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 and zero bias vector. While propagating the prior information through the lower nodes of the tree, τ_b is the scale factor determining the influence of the VTLN prior on the CSMAPLR adaptation technique. The value of the scale factor can be empirically estimated depending on the availability of adaptation data. Scale factors in the range of 1 to 10000 are used to generate adaptation transforms and objective (MCD) score is used as the metric to determine the apt value.

The characteristics estimated by VTLN when propagated to the nodes of the tree structure are expected to improve the speaker specific transform estimation for CSMAPLR. More specifically, VTLN has been shown to be closer to the average voice, and hence better in naturalness [3] and CSMAPLR is known to bring in better speaker similarity when very little adaptation data is available. A-priori, combination of these two is expected to give improved performance with respect to naturalness and speaker similarity.

3. EVALUATIONS WITH VTLN AS PRIOR

3.1. Experimental Setup

The HMM speech synthesis system (HTS) [1] is used for generating the statistical 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 is used to synthesize speech from the parameters generated using HTS. The HMM topology is five-state and left-to-right with no skip states. Speech features are 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 is built using a UK English speech corpus including 5 hours of clean speech data uttered by an RP professional narrator (RJS). The evaluation experiments are performed on another UK English test speaker (Roger). Subjective listening tests are performed by 11 subjects using the Blizzard challenge 2010 test sentences for naturalness, speaker similarity and intelligibility with different amounts of adaptation data and different values of the scale factor.

The subjective tests are based on mean opinion scores (MOS) of naturalness and ABX scores for speaker similarity. The synthesized utterances are rated on a 5-point scale, 5 being "completely natural" and 1 being "completely unnatural". The model (speaker used to train the model) and the target speaker are given as the two reference speakers in the ABX test for finding speaker similarity. Only the spectral stream is transformed with different adaptation techniques; other streams (logF0, bndap and duration) are unadapted or the same as generated for the speaker used to train the model. The subjective evaluations are also performed for intelligibility using semantically unpredictable sentences where subjects listen to the speech utterances and are asked to type the corresponding text. The score for intelligibility is based on the word error rate (WER) for the text entered by the listeners. In addition, objective evaluation based on the mel-cepstral distance (MCD) was also 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 objective evaluations for the test speaker.

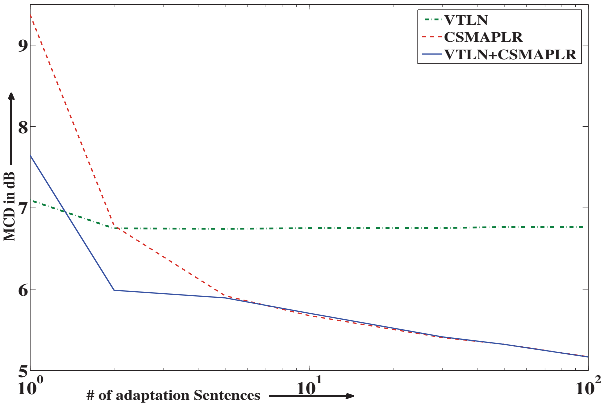
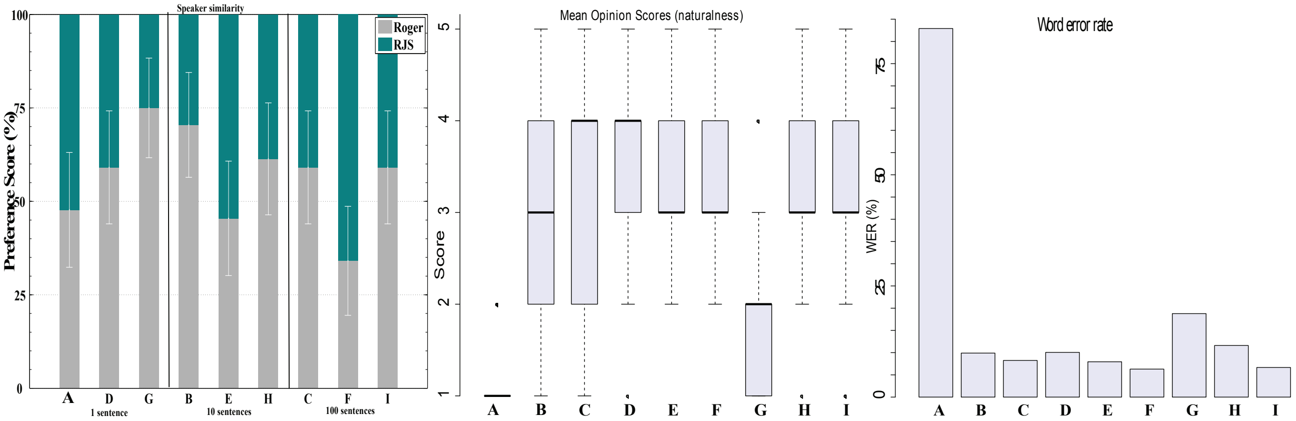


Fig. 1: MCD for VTLN, CSMAPLR and the proposed VTLN-CSMAPLR.

3.2. Results and Discussion

The values of the MCD scores for different amounts of adaptation data are plotted in the Figure 1. The figure shows the MCD score for the scale factor (τ_b) of 1000 (which was empirically determined to be appropriate) for both CSMAPLR and VTLN+CSMAPLR. 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 for the adaptation and that 2) the CSMAPLR improves the MCD to around 6dB 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 VTLN-CSMAPLR 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 VTLN-CSMAPLR technique outperforms the VTLN technique; its distortion is around 6dB.

The listening tests were performed with 1, 10 and 100 adaptation sentences. The evaluation results of the listening tests are shown in Figure 2. From the speaker similarity results, we can see that VTLN works best when the number of adaptation sentences is one and also that VTLN-CSMAPLR outperforms CSMAPLR with one adaptation sentence. There is no significant difference among the CSMAPLR and VTLN-CSMAPLR adaptation methods with 10 or 100 adaptation sentences, both outperforms the VTLN adaptation. From the results on naturalness, we see that VTLN does not improve naturalness even if more data is used. However, VTLN and VTLN-CSMAPLR both give better results than CSMAPLR with one adaptation sentence. From the intelligibility evaluation, we observe that there is no significant difference between VTLN and VTLN-CSMAPLR with 1, 10 and 100 sentences, but, on the other hand, we can see that CSMAPLR has significantly degraded intelligibility with one adaptation sentence. In these results, VTLN with single adaptation sentence is preferred even for speaker similarity only because the test speaker is very close to the speaker used to generate the speaker dependent model (both are RP English male speakers) and VTLN was much better in naturalness. With target speakers very dif-



System symbol Correspondence

System	Adaptation	Adp #
A	CSMAPLR	1
B	CSMAPLR	10
C	CSMAPLR	100
D	VTLN	1
E	VTLN	10
F	VTLN	100
G	VTLN-CSMAPLR	1
H	VTLN-CSMAPLR	10
I	VTLN-CSMAPLR	100

Significance at 1% level for **similarity**

	A	B	C	D	E	F	G	H	I
A	-	0	0	0	0	0	1	0	0
B	0	-	0	0	0	1	0	0	0
C	0	0	-	0	0	0	0	0	0
D	0	0	0	-	0	0	0	0	0
E	0	0	0	0	-	0	1	0	0
F	0	1	0	0	0	-	1	0	0
G	1	0	0	0	1	1	-	0	0
H	0	0	0	0	0	0	0	-	0
I	0	0	0	0	0	0	0	0	-

Significance at 1% level for **naturalness**

	A	B	C	D	E	F	G	H	I
A	-	1	1	1	1	1	1	1	1
B	1	-	0	0	0	0	1	0	0
C	1	0	-	0	0	0	1	0	0
D	1	0	0	-	0	0	1	0	0
E	1	0	0	0	-	0	1	0	0
F	1	0	0	0	0	-	1	0	0
G	1	1	1	1	1	1	-	1	1
H	1	0	0	0	0	0	1	-	0
I	1	0	0	0	0	0	1	0	-

Significance at 1% level for **WER**

	A	B	C	D	E	F	G	H	I
A	-	1	1	1	1	1	1	1	1
B	1	-	0	0	0	0	0	0	0
C	1	0	-	0	0	0	0	0	0
D	1	0	0	-	0	0	0	0	0
E	1	0	0	0	-	0	0	0	0
F	1	0	0	0	0	-	0	0	0
G	1	0	0	0	0	0	-	0	0
H	1	0	0	0	0	0	0	-	0
I	1	0	0	0	0	0	0	0	-

Fig. 2: Listening tests results. There are three columns of plots and tables which are, from left to right, similarity to original speaker, mean opinion score for naturalness, and intelligibility. The similarity is an ABX plot with whiskers for 95% confidence interval. Here systems are permuted differently for readability. Naturalness plot on the upper row is a box plot where the median is represented by a solid bar across a box showing the quartiles and whiskers extend to 1.5 times the inter-quartile range. The system-symbol correspondence is shown in the first table in the bottom row. The rest of the tables in the bottom row indicate significant differences between pairs of systems, based on Wilcoxon signed rank tests with alpha Bonferoni correction (1% level); '1' indicates a significant difference.

ferent from the model speaker, the speaker similarity of VTLN will be very poor compared to that of VTLN-CSMAPLR or CSMAPLR.

4. CONCLUSIONS

We conclude that the VTLN prior can significantly improve the CSMAPLR adaptation performance when the adaptation data is very limited and unlike VTLN, can scale up to the performance of CSMAPLR with more adaptation data. This paper has presented a novel idea for combining the merits of CSMAPLR and VTLN adaptation, resulting in an improved adaptation technique. An efficient algorithm was presented to use the VTLN transformation matrix as prior information for the existing CSMAPLR adaptation. Performance improvements were shown, especially when very little adaptation data was available. The future work is to perform more detailed evaluations in different scenarios and to use multiple VTLN transforms as priors for different phoneme classes instead of a single VTLN transform at the root node.

5. 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, P. N. Garner, and H. Liang, "Implementation of VTLN for statistical speech synthesis," in *Proc. of the 7th ISCA Speech Synthesis Workshop*, Kyoto, Japan, Sept. 2010, pp. 224–229.

[4] J. W. McDonough, *Speaker Compensation with All-Pass Transforms*, Ph.D. thesis, John Hopkins University, 2000.

[5] 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*, South Korea, 2004, pp. 1953–1956.

[6] 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.

[7] 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.

[8] 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.

[9] C. Breslin, K.K. Chin, M.J.F. Gales, K. Knill, and H. Xu, "Prior information for rapid speaker adaptation," in *Proc. of Interspeech*, Japan, 2010, pp. 1644–1647.

[10] 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.

[11] W. Chou, "Maximum a posterior linear regression with elliptically symmetric matrix variate priors," in *Proc. of Eur. Conf. Speech Communication Technology*, Budapest, Hungary, 1999.

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

[13] 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.

[14] 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.

[15] K. Shinoda and C. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Transactions on Speech Audio Processing*, vol. 9, pp. 276–287, Mar. 2001.