IDIAP RESEARCH REPORT

BIAS ADAPTATION FOR VOCAL TRACT LENGTH NORMALIZATION

Lakshmi Saheer Junichi Yamagishi
Philip N. Garner John Dines

Idiap-RR-12-2013

APRIL 2013

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

Bias Adaptation for Vocal Tract Length Normalization

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

[1] Idiap Research Institute, Martigny, Switzerland
[2] École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[3] Centre for Speech Technology Research, University of Edinburgh, U.K.
{lsaheer, pgarner, dines}@idiap.ch, jyamagis@inf.ed.ac.uk

Abstract

Vocal tract length normalisation (VTLN) is a well known rapid adaptation technique. VTLN as a linear transformation in the cepstral domain results in the scaling and translation factors. The warping factor represents the spectral scaling parameter. While, the translation factor represented by bias term captures more speaker characteristics especially in a rapid adaptation framework without having the risk of over-fitting. This paper presents a complete and comprehensible derivation of the bias transformation for VTLN and implements it in a unified framework for statistical parametric speech synthesis and recognition. The recognition experiments show that bias term improves the rapid adaptation performance and gives additional performance over the cepstral mean normalisation factor. It was observed from the synthesis results that VTLN bias term did not have much effect in combination with model adaptation techniques that already have a bias transformation incorporated.

Index Terms: Vocal Tract Length Normalization, Bias term, Hidden Markov models, Speech synthesis, Speech recognition

1. Introduction

Recent advances in the field of statistical parametric speech synthesis [1] has opened up new portfolio of applications like the personalized speech-to-speech (S2S) translation system. An unified framework of hidden Markov model (HMM) based speech synthesis and recognition is very useful for common modelling and speaker adaptation techniques across the S2S system. Such systems demand that the speaker characteristics be embedded into the output speech from the very first input from the user. This demands for rapid speaker adaptation techniques which require very little adaptation data. An unified framework also allows the same adaptation techniques be used across both automatic speech recognition (ASR) and text-to-speech (TTS) synthesis system.

Vocal tract length normalization (VTLN) is a rapid speaker adaptation technique that has been successfully implemented in both statistical speech synthesis [2] and speech recognition [3]. The vocal tract length (VTL) is inversely proportional to the formant frequencies and VTLN normalizes the cepstral features to represent an average VTL. The same technique can adapt the average voice to a particular target speaker. Similar to other linear transformations, VTLN could be implemented as a linear transformation of the cepstrum or equivalently model parameters. The linear transformations usually have two important terms representing translation and scaling. The warping factor in a VTLN implementation represents the scaling. But, the translation term usually referred to as bias is often ignored mainly for the convenience of performing the warping in the speech spectral domain.

VTLN having very limited parameters (a single parameter in the case of bilinear transform) is extremely useful as a rapid adaptation technique. But, at the same time, a single parameter limits the number of speaker characteristics that can be captured by the adaptation technique. It was shown earlier by the authors that VTLN is a powerful rapid adaptation technique and further that it can be effectively combined with other linear transformation techniques like constrained structural maximum a posteriori linear regression (CSMAPLR) [4, 5] to improve rapid adaptation performance for both ASR and TTS. But, these other linear transformations have a multitude of parameters to be estimated. As a trade-off between capturing effective speaker characteristics for rapid adaptation performance and limiting the number of the parameters to be estimated, it might be fruitful to implement VTLN along with the bias term. In this case, the extra parameters in the bias terms could capture more speaker characteristics and still performing as a rapid adaptation technique since the number of parameters in the bias term is limited (usually of the order of the feature dimensionality).

Bias is a very important term in adaptation using linear transformations and influences the performance a lot. It was shown in [6] that a set of offset transforms alone without any scaling/rotation for the mean of the Gaussian models, termed as shift-MLLR, could be used for generating better speaker adaptive models. The offset terms are the bias terms of the speaker transformation. Current implementations of VTLN do not estimate a bias term. This paper presents a complete derivation of the bias term for VTLN. Unlike the scaling matrix A, derivation of bias is very different for other maximum likelihood linear transformations (MLLTs) and VTLN. There has been earlier attempts to use the bias term from MLLTs with VTLN which resulted in performance improvements in an ASR system [7, 8]. This paper presents a complete derivation of the bias term for VTLN and evaluates this term in order to validate the hypothesis that it is an important factor in any linear adaptation technique.

Bias term achieves the translation of the cepstra and can be considered to be equivalent to the cepstral mean and variance normalisation (CMVN). In CMVN, robustness against additive noise is achieved by linearly transforming the cepstral trajectories to have zero mean and unit variance. This works investigates if the bias term is just equivalent to the CMVN and if in deed can provide additional performance improvements compared to CMVN. From our earlier studies and also discussions with other research groups, CMVN was observed to be not performing very well in a TTS system. This paper investigates if this is true for the case of bias transformation as well. Experiments are performed in an unified framework for both ASR and TTS. The paper is organised as follows. The derivation of the

bias term in VTLN is presented in the next section, followed by experimental evaluations on both ASR and TTS systems in section 3 and conclusions in section 4.

2. Bias for VTLN

VTLN involves three main parameters, (1) warping factor(s), (2) a warping function, and (3) an optimization criterion. The mel-generalized cepstral (MGCEP) [9] features are commonly used in the statistical parametric speech synthesis and bilinear transform based VTLN can be shown to be represented by a linear transformation of these features [2]. MGCEP features are very different from the conventional mel-frequency cepstral coefficients (MFCC) used in an ASR system. MGCEP features have a cepstral domain warping instead of a mel filter bank based frequency warping resulting in an invertible set of features between the cepstral and spectral domain. This property makes them particularly attractive in generative models like the statistical parametric speech synthesis systems and also provides a basis for a more accurate cepstral domain linear transformation framework for VTLN which is usually represented by a spectral transformation and only approximated as a equivalent cepstral transform.

Implementing VTLN as a linear transformation in the cepstral domain involves two main components, viz. translation and scaling represented as $\mathbf{W} = [\mathbf{A}, \mathbf{b}]$. Where, A represents the scaling represented by the warping matrix [10] and b represents translation usually referred to as bias. Current implementations of VTLN tend to ignore the bias term for the sake of convenience of representing VTLN as a spectral transformation. But, it can be postulated that the bias term can bring more speaker characteristics and still being limited in the number of parameters can give performance improvements in a rapid adaptation framework. The derivation of the bias term for VTLN is shown below. The auxiliary function to optimize the VTLN based feature adaptation using the maximum likelihood (ML) technique is given by:

$$Q = \log \mathcal{L}(\mathbf{x}(t); \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{A}_\alpha, \mathbf{b}) \tag{1}$$

where, $\mathcal{L}$ represents the likelihood function, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ represent the mean and variance of the model with $\mathbf{x}$ as the feature vector. $\mathbf{A}_\alpha$ represents the VTLN transformation matrix with the warping factor α and $\mathbf{b}$, the bias term for VTLN to be derived.

Using a mixture of Gaussians for the state probability distributions with γ_m as the occupancy for each mixture, m yields the following equation with a Jacobian term given by $\log|\mathbf{A}_\alpha|$.

$$Q = \sum_t \sum_m \gamma_m \Big[\log\left(N(\mathbf{A}_\alpha x_t + b \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)\right) + \log|\mathbf{A}_\alpha| \Big] \tag{2}$$

$$= \sum_t \sum_m \gamma_m \Big[\log\Big(\frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma_m}|}} \exp\Big\{ -\frac{1}{2}(\mathbf{A}_\alpha \mathbf{x}_t + \mathbf{b} - \boldsymbol{\mu}_m)^T {\boldsymbol{\Sigma}_m}^{-1} (\mathbf{A}_\alpha \mathbf{x}_t + \mathbf{b} - \boldsymbol{\mu}_m) \Big\} \Big) + \log|\mathbf{A}_\alpha| \Big] \tag{3}$$

Applying log and ignoring constants,

$$Q = \sum_t \sum_m \gamma_m \Big[-\frac{1}{2}\log|\boldsymbol{\Sigma}_m| - \frac{1}{2}(\mathbf{A}_\alpha \mathbf{x}_t + \mathbf{b} - \boldsymbol{\mu}_m)^T {\boldsymbol{\Sigma}_m}^{-1} (\mathbf{A}_\alpha \mathbf{x}_t + \mathbf{b} - \boldsymbol{\mu}_m) + \log|\mathbf{A}_\alpha| \Big] \tag{4}$$

Ignoring the terms independent of b results in

$$Q = \sum_t \sum_m \gamma_m \Big[-\frac{1}{2}(\mathbf{A}_\alpha \mathbf{x}_t + \mathbf{b} - \boldsymbol{\mu}_m)^T {\boldsymbol{\Sigma}_m}^{-1} (\mathbf{A}_\alpha \mathbf{x}_t + \mathbf{b} - \boldsymbol{\mu}_m) \Big] \tag{5}$$

To optimize this auxiliary function using expectation maximization (EM), calculate the derivative of this function on bias 'b'. Following a similar derivation for mean in [11] and [12].

$$Q = -\frac{1}{2} \sum_t \sum_m \gamma_m \Big[(\mathbf{A}_\alpha \mathbf{x}_t + \mathbf{b} - \boldsymbol{\mu}_m)^T {\boldsymbol{\Sigma}_m}^{-1} (\mathbf{A}_\alpha \mathbf{x}_t + \mathbf{b} - \boldsymbol{\mu}_m) \Big] \tag{6}$$

Using the standard matrix quadratic differential calculus formula:

$$\frac{\partial}{\partial x}(\boldsymbol{y} - \mathbf{x})^T \mathbf{A}(\boldsymbol{y} - \mathbf{x}) = -\mathbf{A}(\boldsymbol{y} - \mathbf{x}) - \mathbf{A}^T(\boldsymbol{y} - \mathbf{x}) \tag{7}$$

$$\frac{\partial Q}{\partial b} = \frac{1}{2} \sum_t \sum_m \gamma_m \Big[\boldsymbol{\Sigma}_m^{-1} (\mathbf{A}_\alpha \mathbf{x}_t - \boldsymbol{\mu}_m + \mathbf{b}) + {\boldsymbol{\Sigma}_m^{-1}}^T (\mathbf{A}_\alpha \mathbf{x}_t - \boldsymbol{\mu}_m + \mathbf{b}) \Big] \tag{8}$$

$$= \sum_t \sum_m \gamma_m \boldsymbol{\Sigma}_m^{-1} (\mathbf{A}_\alpha \mathbf{x}_t - \boldsymbol{\mu}_m + \mathbf{b}) \tag{9}$$

Equating the RHS to zero to find the maximum

$$\frac{\partial Q}{\partial b} = \sum_t \sum_m \gamma_m \boldsymbol{\Sigma}_m^{-1} (\mathbf{A}_\alpha \mathbf{x}_t - \boldsymbol{\mu}_m + \mathbf{b}) = 0 \tag{10}$$

$$-\sum_t \sum_m \gamma_m \boldsymbol{\Sigma}_m^{-1} (\mathbf{A}_\alpha \mathbf{x}_t - \boldsymbol{\mu}_m) = \sum_t \sum_m \gamma_m \boldsymbol{\Sigma}_m^{-1} \mathbf{b} \tag{11}$$

Multiplying both sides by the inverse of the statistics over the inverse of the covariance ($\boldsymbol{\Sigma}_m^{-1}$):

$$b = -\Big(\sum_t \sum_m \gamma_m \boldsymbol{\Sigma}_m^{-1} \Big)^{-1} \sum_t \sum_m \gamma_m \boldsymbol{\Sigma}_m^{-1} (\mathbf{A}_\alpha \mathbf{x}_t - \boldsymbol{\mu}_m) \tag{12}$$

Using a diagonal covariance matrix:

$$b = -\Big(\sum_t \sum_m \gamma_m \sum_i \frac{1}{\boldsymbol{\sigma}_{m,i}^2} \Big)^{-1} \times \sum_t \sum_m \gamma_m \sum_i \frac{(\mathbf{A}_{\alpha,i} \mathbf{x}_{t,i} - \boldsymbol{\mu}_{m,i})}{\boldsymbol{\sigma}_{m,i}^2} \tag{13}$$

The resulting bias term could be implemented with the linear VTLN transformation. Since it is not possible to estimate the transformation matrix and bias term simultaneously, a better approach would be to iteratively optimize the two components each estimated in alternative iterations considering the effect of the other term. The bias term estimation is implemented in the HMM-based speech toolkit (HTK) and evaluated with both ASR and TTS experiments.

3. Experiments & Results

This section presents both ASR and TTS experiments in an unified framework. The motivation of these experiments is to test the hypothesis that bias term is useful in improving the rapid adaptation performance of VTLN. Additionally, in the case of

# of adp sent	CSMAPLR	VTLN	Bias	VTLN+Bias
2	15.19	39.12	16.52	16.36
5	13.18	39.05	16.05	16.22
10	12.75	38.89	16.13	16.12
40	11.9	38.61	16.25	16.02

Table 1: WER for different adaptation techniques. The SAT model without any adaptation on the test data gives a WER of 40.31%

an ASR system, it is interesting to know if bias can provide further improvements to a CMVN system. In the case of TTS, the best rapid adaptation system is the combination of VTLN as a prior to the CSMAPLR system. The experiments are performed on TTS to determine if the bias term has any influence as a prior on the CSMAPLR adaptation. Different methods are presented to combine VTLN and bias transformation as prior to the CSMAPLR adaptation. The main aim of the TTS experiments is to find the best method to combine these transformations in a rapid adaptation framework.

3.1. ASR

Following the work in [13], this section presents ASR experiments similar to the set-up in a TTS system. It is to show that the techniques developed in this research can in fact be used for both TTS and ASR, and the experiments are in accordance to the unification theme of a speech-to-speech translation system. It should be noted that the ASR results are based on the experiments in [13] and does not represent the state of the art for this corpus since the idea is to use a common features and modeling techniques for ASR and TTS.

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 [13][1]. SAT models were generated using the (American English) Wall street journal (WSJ0) database. The evaluations were carried out using Spoke 4 (S4) task of the November 1993 CSR evaluations (same as the ones used in the baseline system mentioned above). The adaptation was carried out off-line using the rapid enrollment data (for condition C3) which comprises 40 adaptation utterances for each of the 4 speakers. The system uses 5k word list with a bigram language model.

ASR performance is presented for the bias parameter of VTLN. The bias parameter estimation is independent from that of the warping factor for VTLN An efficient method should be devised to combine these two components in an effective manner. For these experiments VTLN warping factor ($[\mathbf{A}_\alpha,0]$) was estimated initially and then this transformation was used as a parent transform to estimate the bias term ($[\boldsymbol{I}, \mathbf{b}]$ where $\boldsymbol{I}$ is the identity matrix). Finally, the two components were combined into a single transformation matrix ($\mathbf{A}_\alpha$,$\mathbf{b}$]). Also, a bias alone transformation was estimated using [$\boldsymbol{I}$,0] as the parent transformation. ASR evaluations were performed using all the schemes.

The results are presented in Table 1. The table shows word error rates (WER) for different amounts of adaptation data ranging from two to 40 adaptation sentences comparing four different systems: (1) CSMAPLR system, (2) VTLN adaptation ($[\mathbf{A}_\alpha,0]$) representing a single warping factor, (3) bias alone transformation ($[\boldsymbol{I}, 0]$) denoted as Bias and (4) VTLN with bias ($[\mathbf{A}_\alpha,\mathbf{b}]$) using the method mentioned above denoted as VTLN+Bias The results show that bias is an important term in the VTLN transform estimation. Just with the bias factor, the recognition performance improves a lot and becomes comparable to the CSMAPLR system unlike the VTLN without bias adaptation. The overall results are still not better than the CSMAPLR system since the number of parameters in the transformations (VTLN plus bias) are limited compared to the CSMAPLR transformations. The idea of this paper is not to directly compare the performances of VTLN+Bias with CSMAPLR technique or present VTLN as a superior adaptation technique compared to CSMAPLR. The idea is to understand if the majority of the performance gains achieved by the linear transformation systems like CSMAPLR can be attributed to the bias term in these transformations.

Experiments are also performed to test the influence of bias term in the presence of CMVN compensation. Two speaker adaptive trained (SAT) models were trained, one with and other one without using the CMVN. Then experiments are performed on each of these models with bias adaptation. In the case of CMVN model, CMVN is also used on the test data. The results as WER are summarised in Table 2.

# of adp sent	Using CMVN	No CMVN
SAT model	35.9	40.31
2	16.16	16.52
5	16.09	16.05
10	16.18	16.13
40	16.16	16.25

Table 2: WER for bias term of VTLN. SAT model does not use any adaptation on the test data.

The results show that the SAT model (without any adaptation on test data) performance improves to 35.9% from 40.31% when CMVN is used. The bias transformation further reduces the WER to around 16%. But, the results with bias term does not change with or without CMVN. This suggests that bias transformation includes CMVN and is not limited to just the cepstral normalisation. The additional performance improvement can be obtained from the bias transformation when compared to CMVN. Additionally, CMVN can be ignored if the system uses a bias transformation.

3.2. Gender Transforms in TTS

A mismatch scenario where VTLN can perform well is the wide variation in the speakers used in training and adaptation. This variation might be because speakers are from different genders for training and adaptation. The hypothesis is that VTLN can represent the differences in vocal tract length across gender and prove to be beneficial in such scenarios. For this purpose, we have used the CSTR VCTK corpus[2]. 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

[1]The baseline system is the system 'd' in Table IX of [13], which has 13% word error rate (WER). The baseline system reported in [13] 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%.

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

Techniques	Male Test		Female Test	
	One sentence	five sentences	One sentence	five sentences
VTLN+Bias	6.0952	6.0332	6.0197	5.9827
VTLN-CMAPLR	5.9525	5.6150	5.8285	5.7947
BiasCN-VTLN-CMAPLR	6.0977	5.7830	5.8999	5.6304
VTLN+Bias-CMAPLR	5.9795	5.7950	5.8653	5.6148

Table 3: MCD for gender dependent female models using bias for VTLN prior

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 female gender-dependent average voice models to them to see the impact of the VTLN with bias as prior from many speakers, especially in cross-gender cases.

The HMM speech synthesis system (HTS) [14] 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 [15] 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. 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.

Evaluations were performed to check the influence of bias on the rapid adaptation. Bias has to be combined with VTLN and the CSMAPLR transformations to achieve the best rapid adaptation performance. Two methods were adopted to this end. Bias can be seen as cepstral shift similar to cepstral mean normalization. Hence, the first method involves using bias as a cepstral normalization technique for the model means and then continue as in [4] using VTLN as prior for CSMAPLR transformation. Second method involves estimating VTLN and bias iteratively with one as a parent transform of the other and finally combining them both into a single transformation. This combined VTLN and bias transformation can act as the prior for CSMAPLR. The hypothesis here is that bias term can add improvements to rapid adaptation when combined with VTLN and CSMAPLR. Objective evaluations as MCD scores were estimated on the gender dependent female models. Both male and female test speakers were evaluated. The results compare four different systems:

1. VTLN and bias estimated iteratively with each one as a parent transform of the other. This transformation (referred to as VTLN+Bias) is used to adapt the model.
2. VTLN is used as a prior to the root node of the CSMAPLR transformation (referred to as VTLN-CSMAPLR). This is the same system presented in our earlier work [4].
3. Bias used as a cepstral mean normalization and then, VTLN transformations estimated and used as prior at the root node of the CSMAPLR transformation. This system is named BiasCN-VTLN-CSMAPLR
4. VTLN along with Bias as presented in the case 1 being used as a prior at the root node of CSMAPLR transformation (referred to as VTLN+Bias-CSMAPLR)

Randomly chosen single and five adaptation sentences were used to generate the transforms for each method. 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 objective measure. The results are shown in Table 3 for one and five adaptation sentences. The results show no perceivable difference when bias is combined using the techniques proposed. The hypothesis cannot be established that the VTLN with bias acts as a better prior for CSMAPLR adaptation. This could be because when acting just as a prior, bias term is not able to contribute to performance enhancement. The combination of bias with VTLN and CSMAPLR requires further investigation in order to utilize full potential of the bias term.

4. Conclusions

This paper presented the derivation of the bias term for VTLN and implemented it successfully in the HTS system to perform both ASR and TTS experiments. The ASR experiments validated the fact that bias transformation is important especially when the amount of adaptation data is limited and achieves performance comparable to the CSMAPLR transformation. The idea of the paper is not to compare the VTLN or Bias transformation to powerful model adaptation techniques like CSMAPLR. But, the results show that the majority of the performance gain achieved by the linear transformation like CSMAPLR could be just due to the presence of a bias factor. This supports the earlier observations in [6]. Bias gives performance improvements even in the absence of VTLN warping and hence seems to be a good trade-off between speaker characteristics captured with limited amount of adaptation data and the number of parameters to be estimated for adaptation. From the CMVN experiments it was established that bias transformation includes cepstral mean normalisation but, is not limited to this. Hence, if bias transformation is used in adaptation to a target speaker, CMVN could be ignored saving some computational overhead.

As mentioned earlier, CMVN is not known to give additional improvements in the case of TTS. From the experiments presented in the paper, the same is true for the bias transformation for TTS. Different methods were presented to combine VTLN, bias and CSMAPLR transformations including using bias as a cepstral mean normalisation technique. In TTS, the bias term does not seem to have a great influence especially when combined with powerful model based adaptation technique as a prior information or as a cepstral mean normalisation term. This requires further study to clarify what other contributions can be made by the bias term in statistical parametric speech synthesis.

5. Acknowledgements

Authors would like to thank the HASLER funding for V-FAST project for this work. The current work is funded by the D-Box project.

6. 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] 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.

[3] 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.

[4] L. Saheer, J. Yamagishi, P. N. Garner, and J. Dines, "Combining vocal tract length normalization with hierarchial linear transformations," in *Proc. of ICASSP*. Kyoto, Japan: IEEE SPS, Mar. 2012, pp. 4493–4496.

[5] 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.

[6] J. Lööf, C. Gollan, and H. Ney, "Speaker adaptive training using shift-mllr," in *Proceedings of Interspeech*, Brisbane, Australia, Sep. 2008, pp. 1701–1704.

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

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

[9] 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.

[10] 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.

[11] P. N. Garner and W. J. Holmes, "On the robust incorporation of formant features into hidden Markov models for automatic speech recognition," in *Proc. of ICASSP*, vol. 1, 1998, pp. 1–4.

[12] L. A. Liporace, "Maximum-likelihood estimation for multivariate observations of markov sources," *IEEE Transactions on Information Theory*, vol. 28, no. 5, pp. 729–734, 1982.

[13] 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.

[14] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[15] 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.