VTLN-BASED RAPID CROSS-LINGUAL ADAPTATION FOR STATISTICAL PARAMETRIC SPEECH SYNTHESIS

Lakshmi Saheer Hui Liang John Dines
Philip N. Garner

Idiap-RR-12-2012

APRIL 2012

VTLN-Based Rapid Cross-Lingual Adaptation for Statistical Parametric Speech Synthesis

Lakshmi Saheer [+#1], Hui Liang [+#2], John Dines [#3], Philip N. Garner [#4]

[#] *Idiap Research Institute, Martigny, Switzerland*
[+] *École Polytechnique Fédérale de Lausanne (EPFL), Switzerland*
[1] `lsaheer@idiap.ch`, [2] `hliang@idiap.ch`, [3] `dines@idiap.ch`, [4] `pgarner@idiap.ch`

***Abstract*—Cross-lingual speaker adaptation (CLSA) has emerged as a new challenge in statistical parametric speech synthesis, with specific application to speech-to-speech translation. Recent research has shown that reasonable speaker similarity can be achieved in CLSA using maximum likelihood linear transformation of model parameters, but this method also has weaknesses due to the inherent mismatch caused by differing phonetic inventories of languages. In this paper, we propose that fast and effective CLSA can be made using vocal tract length normalization (VTLN), where strong constraints of the vocal tract warping function may actually help to avoid the most severe effects of the aforementioned mismatch. VTLN has a single parameter that warps spectrum. Using shifted or adapted pitch, VTLN can still achieve reasonable speaker similarity. We present our approach, VTLN-based CLSA, and evaluation results that support our proposal under the limitation that the voice identity and speaking style of a target speaker don't diverge too far from that of the average voice model.**

***Index Terms*—vocal tract length normalization, cross-lingual speaker adaptation, rapid speaker adaptation, HMM-based speech synthesis**

I. Introduction

The ability to transform voice identity in text-to-speech synthesis (TTS) has been an important area of research with applications in medical, security and entertainment industries. One specific application that has seen considerable interest by the speech research community is that of speech-to-speech translation, where the challenge of voice transformation is further compounded by the differing languages of target speaker data and output synthesis. Statistical parametric synthesis [1] has proven to be a particularly flexible and robust framework for voice transformation, leveraging off a range of speaker adaptation techniques previously developed for automatic speech recognition (ASR). The extension of these approaches to a cross-lingual setting is commonly referred to as cross-lingual speaker adaptation (CLSA).

CLSA takes speech data in one language and uses this to adapt a set of acoustic models for synthesis in a different language. Unlike in intra-lingual speaker adaptation, it is evident that the correspondence between adaptation data and the acoustic models to be adapted is largely lost at the linguistic level. To date, the most successful approaches have relied on the construction of a set of mapping rules between acoustic model distributions (i.e. HMM states) for the two languages, thus establishing sub-phonemic (or senone-level) correspondence between the two languages [2]. Given this state mapping, CLSA may be performed using conventional speaker adaptation techniques such as constrained structural maximum *a posteriori* linear regression (CSMAPLR).

Despite the progress that has been made in CLSA, it is evident that the state-of-the-art still lags behind intra-lingual speaker adaptation in terms of synthesis performance (i.e. speaker similarity, speech naturalness, etc.). This is in large part due to the fact that the state-level mapping is still unable to fully account for the inherent mismatch between phonetic inventories of different languages [3]. Vocal tract length may be considered to be inherently language independent, hence, we postulate that VTLN may not suffer from such mismatch issues.

The application of VTLN to statistical parametric speech synthesis has previously been shown to be promising for rapid, intra-lingual speaker adaptation [4], revealing that VTLN-synthesis was able to produce naturalness ratings close to average voice and significantly better than model adaptation techniques like CSMAPLR while still improving speaker similarity over the average voice. This paper investigates the use of VTLN for CLSA, especially in the scenario where very little adaptation data is available. A new framework facilitating supervised rapid CLSA is presented, where HMM state mapping is integrated into bilinear transform-based VTLN. We tested the hypothesis that the constrained nature of VTLN transformation might help to alleviate some problems associated with current CLSA approaches. Experiments were performed on the Mandarin-English language pair and VTLN adaptation was compared with CSMAPLR, which is the best-known robust and rapid adaptation technique in synthesis.

II. Framework for VTLN-based CLSA

CLSA remains a challenging task and relevant literature is sparse as the field draws on several disparate concepts, each non-trivial in its own right [5]–[7]. Previous work on CLSA normally employs CSMAPLR or related adaptation techniques. In the context of intra-lingual speaker adaptation, CSMAPLR has proven effective in capturing main speaker characteristics, but its application in a cross-lingual context has met with less success, especially when multiple adaptation transforms are used [3]. By contrast, VTLN has significantly fewer parameters (typically only one parameter is used to modify the vocal tract warping function) and as such the range of speaker characteristics that can be represented is restricted. However, in the cross-lingual scenario, where CSMAPLR is susceptible to learning not only speaker characteristics, but also undesirable language mismatches, VTLN may provide more acceptable results. The fact that CSMAPLR and our VTLN implementation operate on the underlying HMM distributions in the same manner (i.e. as maximum *a posteriori*/likelihood linear feature transformation) provides a good basis for testing this hypothesis.

A. *Vocal Tract Length Normalization*

Vocal tract length (VTL) varies across different speakers (around 18 cm in males to around 13 cm in females). Formant frequency positions are inversely proportional to VTL, thus, a variation of around 25% in formant center frequencies is observed among speakers. It follows that we can normalize feature vectors extracted from speech of different speakers to represent an average vocal tract – so called *vocal tract length normalization* (VTLN).

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 the average VTL.

In ASR, where a mel- or bark-spaced filter bank is used, the warping function tends to be piecewise linear, and is normally applied directly to spectrum prior to application of the filter bank (thereby making direct warping of features impossible except via spectrum interpolation). By contrast, feature extraction for TTS does not rely on filter bank analysis due to the problem this poses for signal reconstruction. Rather, the analysis approach undertaken is mel-generalized cepstrum (MGCEP) [8], which makes use of a bilinear transform to achieve frequency warping. The bilinear transform of a simple first-order all-pass filter with unit gain leads to:

$$\beta_\alpha(\omega) = \arctan \frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha} \tag{1}$$

where α is the warping factor and ω is the frequency being warped. Since MGCEP already includes a bilinear transform as its spectral warping function to approximate the mel auditory scale, a bilinear transform-based VTLN can thus be implemented as a zero-overhead modification of the MGCEP codec [9]. The frequency warping $\beta_\alpha(\omega)$ can be represented as a linear transformation of the cepstral features [10]:

$$\mathbf{C}\,\mathbf{log}(\mathcal{F}[\beta_\alpha(\omega)]) = \mathbf{A}_\alpha \mathbf{C}\,\mathbf{log}(\mathcal{F}[\omega]) \tag{2}$$

where $\mathbf{C}$ is the discrete cosine transformation (DCT) matrix, **log** represents an element-wise logarithmic function, $\mathcal{F}$ represents the element-wise magnitude of discrete Fourier transformation and $\mathbf{A}_\alpha$ is the transformation matrix representing the bilinear transform. The cepstral features are extracted by applying DCT on the log spectrum. The transformed cepstral features ($\mathbf{x}_\alpha$) can thus be represented as

$$\mathbf{x}_\alpha = \mathbf{A}_\alpha \mathbf{x}. \tag{3}$$

The bilinear transformation generates $\mathbf{A}_\alpha$ in a special form as shown below:

$$\begin{bmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{M-1} \\ 0 & 1-\alpha^2 & 2\alpha(1-\alpha^2) & \cdots & M\alpha^{M-1}(1-\alpha^2) \\ 0 & -\alpha(1-\alpha^2) & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & (-\alpha)^{M-1}(1-\alpha^2) & \cdots & \cdots & \cdots \end{bmatrix}$$

The maximum likelihood optimization is [11]:

$$\hat{\alpha}_s = \arg\max_{\alpha} p\left(\mathbf{A}_{\alpha_s}\mathbf{x}_s \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{q}\right) \tag{4}$$

where $\mathbf{x}_s$ represents original feature vectors to be warped with the warping factor α_s for speaker s, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ represent mean vectors and covariance matrices of average voice models, $\mathbf{q}$ represents state sequences of transcriptions and $\hat{\alpha}_s$ represents the best warping factor for speaker s.

1) Estimation of VTLN Warping Factors: VTLN amounts to linear transformation in the cepstral domain [10] and can be implemented as as equivalent model transformation. Such representation enables the use of techniques like expectation maximization (EM) for finding optimal warping factors [12], [13]. The main advantage of using EM over, say, a grid search is that the resulting warping factor estimation is based on a gradient descent technique which provides finer granularity of α values and efficient implementation in time and space. EM can be embedded into HMM training utilizing the same sufficient statistics as in CSMAPLR and the resulting auxiliary function for VTLN is [4]:

$$Q(\alpha) = \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}_i \mathbf{G}_i \boldsymbol{w}_i^\top - 2\boldsymbol{w}_i \mathbf{k}_i^\top) - \phi \log|\mathbf{A}_\alpha|$$

where

$$\mathbf{G}_i = \sum_{m=1}^{M}\frac{1}{\sigma_{m,i}^2}\sum_{f=1}^{F}\gamma_m \mathbf{x}_f \mathbf{x}_f^\top \quad , \quad \phi = \sum_{f=1}^{F}\sum_{m=1}^{M}\gamma_m$$

$$\text{and} \quad \mathbf{k}_i = \sum_{m=1}^{M}\frac{\mu_{m,i}}{\sigma_{m,i}^2}\sum_{f=1}^{F}\gamma_m \mathbf{x}_f^\top .$$

$\boldsymbol{w}_i$ represents the i^{th} row of the transformation matrix $\mathbf{A}_\alpha$ for an input feature vector $\mathbf{x}$. F and M are the total number of frames and mixtures, respectively. γ_m, μ_m and Σ_m are posterior probabilities and parameters of the Gaussian mixture component m. N is the feature dimensionality. Formulation of VTLN as model transformation leverages many techniques that have been developed from linear transform-based adaptation such as multiple transforms via regression classes and use of prior probability distributions via CSMAPLR. The efficient implementation of VTLN using EM, with Brent's search optimization for synthesis, by Saheer *et al.* [4] is used in this work.

The estimation of VTLN warping factors for TTS turns out to be a more complex problem than that for ASR due to the increased feature dimensionality. This results in issues related to numerical stability, amplification of unmodelled correlation, and Jacobian normalization [14] severely reducing the range of warping factors. These issues were dealt by Saheer *et.al*. [4]. Here, the warping factors are estimated using lower order features without Jacobian normalization, which was shown previously to be effective for TTS [4].

B. Integration of VTLN into State Mapping-Based CLSA

We integrate HMM state mapping, which has proven effective for CLSA [2], into bilinear transform-based VTLN. First of all, we define the language in which speech is synthesized as the *output language* and the language of given adaptation utterances from a target speaker as the *input language*. Two monolingual average voice model sets are established in the input and output languages respectively, $\mathbb{S}^{\text{in}} = \left\{S_1^{\text{in}}, S_2^{\text{in}}, \cdots, S_{N^{\text{in}}}^{\text{in}}\right\}$ and $\mathbb{S}^{\text{out}} = \left\{S_1^{\text{out}}, S_2^{\text{out}}, \cdots, S_{N^{\text{out}}}^{\text{out}}\right\}$, where S refers to state distributions. Following this, a set of state mapping rules, $\mathbb{M}(\cdot)$, is constructed such that

$$\mathbb{M}\left(S_i^{\text{in}}\right) = \arg\min_{S_j^{\text{out}} \in \mathbb{S}^{\text{out}}} D_{\text{K-L}}\left(S_i^{\text{in}}, S_j^{\text{out}}\right), \ \forall S_i^{\text{in}} \in \mathbb{S}^{\text{in}} \qquad (5)$$

where $D_{\text{K-L}}(\cdot,\cdot)$ denotes the symmetric Kullback-Leibler divergence between two Gaussian distributions [1].

Wu *et al.* [2] proposed two ways of applying these state mapping rules: data transfer and transform transfer. It has been observed [3], [7] that data transfer is preferred over transform transfer, thus, the work in this paper is based on data transfer and a cross-lingual warping factor $\hat{\alpha}_s$ is estimated as follows, in a similar fashion to Eq. (4) (the intra-lingual version):

$$\hat{\alpha}_s = \arg\max_{\alpha} p\left(\boldsymbol{A}_{\alpha_s}\boldsymbol{x}_s^{\text{in}} \,\middle|\, \boldsymbol{\mu}^{\text{out}}, \boldsymbol{\Sigma}^{\text{out}}, \mathbb{M}(\boldsymbol{q}^{\text{in}})\right) \qquad (6)$$

where $\boldsymbol{x}_s^{\text{in}}$ is acoustic feature vectors of adaptation data of speaker s, $\boldsymbol{q}^{\text{in}}$ consisting of $\left\{S_i^{\text{in}}\right\}$ is the state sequence of $\boldsymbol{x}_s^{\text{in}}$, $\boldsymbol{\mu}^{\text{out}}$ and $\boldsymbol{\Sigma}^{\text{out}}$ are mean vectors and covariance matrices of an average voice in the output language.

Using a greater number of transforms is generally beneficial to the performance of intra-lingual speaker adaptation. Interestingly, Liang *et al.* [3] discovered the fact was just the opposite in CLSA: It was better to estimate only a single global transform for all state emission distributions when using data transfer. This paper also investigates whether this phenomenon will be observed in VTLN-based CLSA.

III. Investigation

The experiments performed in this paper are mainly focussed on testing two hypotheses:

1) As a highly constrained feature transformation, VTLN may perform better than CSMAPLR in a rapid-CLSA scenario where limited adaptation data is available.
2) Multiple transform-based VTLN will also degrade performance in the cross-lingual scenario, as has been previously observed for CSMAPLR.

In this work we used the Mandarin-English language pair, with Mandarin/English being the input/output language. One Mandarin adaptation utterance and its context-dependent labels were used to generate speaker-specific transforms. The techniques compared were global/multiple VTLN transform and global/multiple CSMAPLR transform based adaptation. A global VTLN transform corresponded to a single speaker-specific warping factor applied to an entire model set. Multiple VTLN transforms corresponded to different speaker-specific and phoneme class-dependent warping factors generated from a regression class tree in the usual fashion. Likewise, a global CSMAPLR transform applied to an entire model set and multiple CSMAPLR transforms were regression class-dependent. The prior weighting for the CSMAPLR transforms were adjusted to an empirically determined value[2] of 1000, which has been previously observed to give the best results with a small amount of adaptation data [15].

A. Experimental Setup

Two average voice synthesis models were trained on the SpeeCon (Mandarin, 12.3 hours) and WSJ SI84 (English, 15.0 hours) corpora in the HTS-2007 framework [1]. The HMM topology was five-state (single mixture, multivariate Gaussians) and left-to-right with no skips. Speech features were 39th-order mel-cepstra, log F0, five-dimensional band aperiodicity, and their delta and delta-delta coefficients, extracted from 16kHz recordings with a window shift of 5ms. Detailed evaluations were performed on a pilot corpus recorded in an anechoic studio in University of Edinburgh by a male, native Mandarin speaker uttering Mandarin and reasonably natural English. Only one Mandarin adaptation utterance of 7.71 seconds was used for transform estimation in all cases. In addition, a limited number of systems were selected for further evaluations with one male and three female speakers from an EMIME bilingual (Mandarin-English) corpus [16] recorded in the same anechoic studio. These four speakers were with the least foreign accents in their spoken English amongst all the speakers in the EMIME bilingual corpus, and only a single Mandarin adaptation utterance of similar duration was used for each of them. The above experimental setup is the same as that of Liang *et al.* [7], except for the source of adaptation and evaluation data.

This paper focuses on cross-lingual adaptation of spectrum. The subjective evaluations were based on AB and ABX tests for naturalness and speaker similarity, respectively. Listeners were presented with two speech samples at a time and asked to judge which one sounded more natural or closer to the voice of a reference sample. Mandarin reference samples were presented to the listeners for judging speaker similarity of synthesized speech in English for ABX tests. The listening tests were performed only on selected pairs of systems that could give the most useful insights with respect to our hypotheses.

B. Evaluation Results and Discussions

It is expected that VTLN produces far more natural-sounding speech than CSMAPLR, since the adaptation of a

[1] We assume that state distributions comprise single Gaussian PDFs as is usual for HTS.

[2] The HTK variable `HADAPT:SMAPSIGMA` was set to 1000.

single parameter prevents gross modification of the average voice model, thereby maintaining the better naturalness of the original average voice model [17].

The initial evaluations were conducted with 4 pairs of systems for the male speaker from the pilot bilingual corpus. Each listener evaluated 80 English utterances in total. The results are plotted in Figure 1 with 95% confidence intervals. It is evident from these figures that VTLN is far more natural compared to CSMAPLR, but the ability to achieve good speaker similarity with VTLN alone is limited.

Based on this result and our own observations, we suppose that the effectiveness of VTLN as a speaker adaptation technique for TTS is dependent on the characteristics of a target speaker – some speakers cannot be sufficiently reproduced using VTLN adaptation while others can. To that end, evaluations were performed with the four speakers from the EMIME bilingual corpus. Only two pairs of systems (average voice vs global-VTLN and global-VTLN vs global-CSMAPLR) were compared for these speakers for finding the effectiveness of VTLN as an adaptation technique. Each listener was presented with 20 pairs of sentences for each of the four speakers, judging naturalness and speaker similarity. Results are plotted in Figure 2. Similar trends are observed in these results. Since the training data for estimation of average voice is dominated by male speakers, better results are observed with VTLN for female test speakers. Unlike the previous case, the VTLN system is preferred over CSMAPLR, even for speaker similarity, mainly because of the fact that VTLN-synthesized speech sounded more natural than CSMAPLR. To further elaborate, neither adaptation technique could exactly reproduce a target speaker's voice characteristics with a little adaptation data. Listeners are unable to separate their preference for naturalness from their judgement of speaker similarity. Hence, the listeners preferred more natural-sounding speech. For the same reason, some male speakers could be judged closer to the average voice in speaker similarity since the average voice is male dominant and better in naturalness when compared to VTLN.

It is also worth noting results of perception experiments in [18], which suggest that the correctness of speaker discrimination is only 51%-61% if two speech samples for comparison are in different languages *and* of different speech types (i.e. natural or speaker-adapted). Thus, judgement of speaker similarity in a CLSA context is already a difficult task regardless of the the approach employed. By contrast, the advantages offered by VTLN-based CLSA with respect to naturalness are quite clear, while the approach still maintains gross speaker qualities (e.g. gender, etc.). The results thus confirm the first hypothesis made at the beginning of this section that VTLN performs better compared to CSMAPLR in a rapid CLSA scenario.

Concerning a comparison of global and multiple transform adaptation approaches, it is clear from the subjective evaluation that multiple transforms provide inferior CLSA performance. This is consistent with earlier studies [4], [7] that showed that while multiple transforms improve the performance of intra-lingual speaker adaptation, a degradation in CLSA performance is observed. We also note that, based on subjective evaluation, multiple transform VTLN-based CLSA was more preferable compared to multiple transform CSMAPLR. The second hypothesis presented at the beginning of this section also proves to be correct for both CSMAPLR and VTLN, with VTLN-based adaptation being more preferable in the multiple transform case as well.

IV. Conclusions

This paper presents a new framework for rapid CLSA using VTLN. A single adaptation utterance in Mandarin from a target speaker was used to generate English speech in that speaker's voice. The results of VTLN adaptation were compared with those of CSMAPLR adaptation. It is observed that VTLN provided better naturalness than CSMAPLR, but at the price of reduced speaker similarity. This is especially evident when target speaker characteristics were far from those of the average voice model. The constrained nature of VTLN could provide some subjective improvements to adaptation using multiple transforms, but overall global transformation still proved the most effective.

The results are promising, especially in the sense that the merits of both CSMAPLR and VTLN can potentially be combined, for instance, by using VTLN matrices as prior information for CSMAPLR. This future research direction could result in improved quality of cross-lingually adapted speech with as little as a single utterance of adaptation data.

Acknowledgment

Research leading to the results in this paper was funded by the Seventh Framework Programme (FP7/2007-2013) of European Union under the grant agreement 213845 (the EMIME project).

References

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

[2] Y.-J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis", in *Proc. of Interspeech*, Sep. 2009, pp. 528–531.

[3] H. Liang and J. Dines, "An analysis of language mismatch in HMM state mapping-based cross-lingual speaker adaptation", in *Proc. of Interspeech*, Sep. 2010, pp. 622–625.

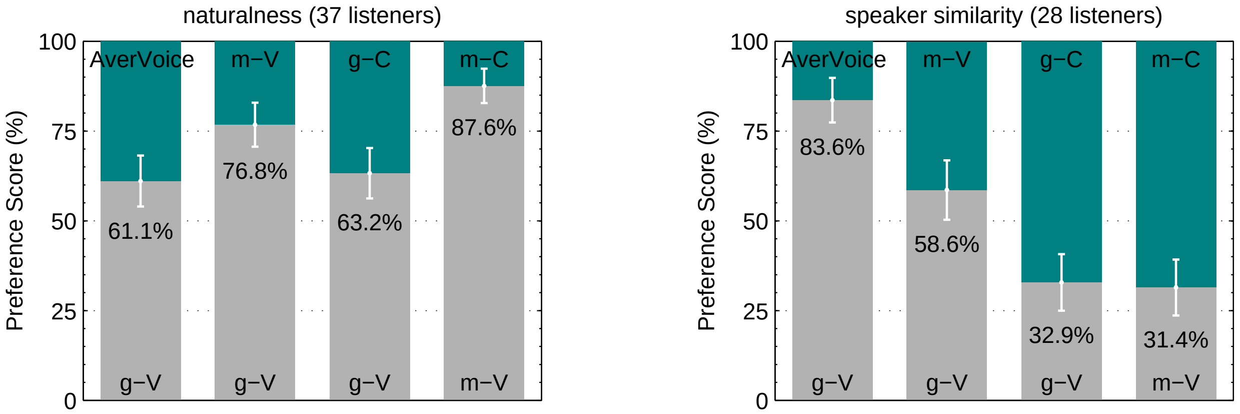
[4] L. Saheer, J. Dines, P. N. Garner, and H. Liang, "Implementation of VTLN for statistical speech synthesis", in *Proc. of the 7th ISCA Speech Synthesis Workshop*, Kyoto, Japan, Sep. 2010, pp. 224–229.

[5] K. Oura, K. Tokuda, J. Yamagishi, S. King, and M. Wester, "Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis", in *Proc. of ICASSP*, Mar. 2010, pp. 4594–4597.

[6] M. Gibson and W. Byrne, "Unsupervised intralingual and cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 895–904, May 2011.

[7] H. Liang, J. Dines, and L. Saheer, "A comparison of supervised and unsupervised cross-lingual speaker adaptation approaches for HMM-based speech synthesis", in *Proc. of ICASSP*, Mar. 2010, pp. 4598–4601.

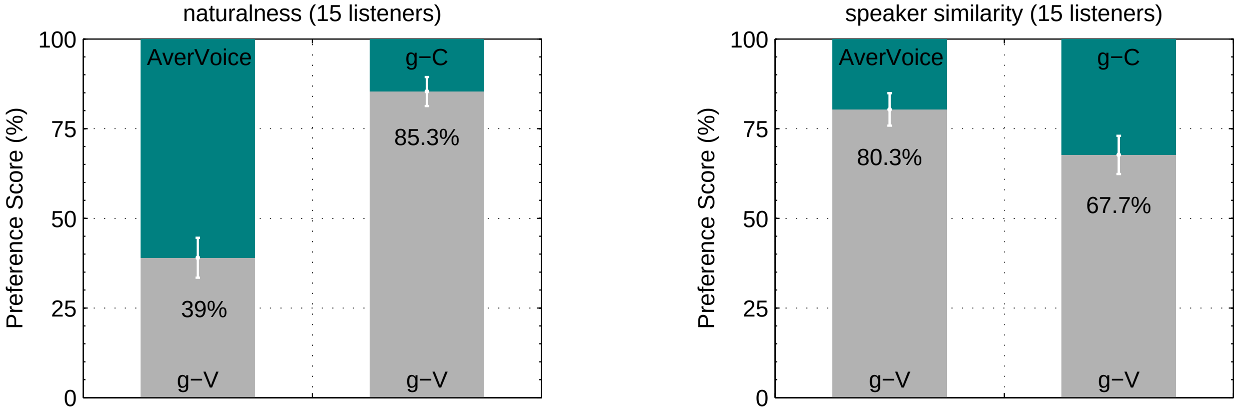
[8] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis – A unified approach to speech spectral estimation", in *Proc. of ICSLP*, vol. 3, Sep. 1994, pp. 1043–1046.



(a) AB test for naturalness

(b) ABX test for speaker similarity

Fig. 1: Results of evaluation on more system combinations for the male speaker in the pilot corpus. The systems are named as (g/m)-(V/C): g/m means *global/multiple*, and V/C means VTLN/CSMAPLR.



(a) AB test for naturalness

(b) ABX test for speaker similarity

Fig. 2: Results for the four target speakers from the EMIME bilingual corpus. The systems are named as (g/m)-(V/C): g/m means *global/multiple*, and V/C means VTLN/CSMAPLR.

[9] 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.

[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] 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.

[12] 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.

[13] P. T. Akhil, S. P. Rath, S. Umesh, and D. R. Sanand, “A computationally efficient approach to warp factor estimation in VTLN using EM algorithm and sufficient statistics”, in *Proc. of Interspeech*, Brisbane, Australia, 2008, pp. 1713–1716.

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

[15] D. Miyamoto, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, “Acoustic compensation methods for body transmitted speech conversion”, in *Proc. of ICASSP*, Apr. 2009, pp. 3901–3904.

[16] M. Wester and H. Liang, “The EMIME Mandarin bilingual database”, University of Edinburgh, U.K., Tech. Rep. EDI-INF-RR1396, Feb. 2011.

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

[18] M. Wester and H. Liang, “Cross-lingual speaker discrimination using natural and synthetic speech”, in *Proc. of Interspeech*, Aug. 2011.