COMBINING VOCAL TRACT LENGTH NORMALIZATION WITH LINEAR TRANSFORMATIONS IN A BAYESIAN FRAMEWORK

Lakshmi Saheer   Junichi Yamagishi
Philip N. Garner   John Dines

Combining Vocal Tract Length Normalization with Linear Transformations in a Bayesian Framework

Lakshmi Saheer [+#1], Junichi Yamagishi [*2], Philip N. Garner [#3], John Dines [#4]

[+] *Ecole Polytechnique Federale de Lausanne, Switzerland*
[#] *Idiap Research Institute, Switzerland*
[1] lsaheer@idiap.ch [3] pgarner@idiap.ch [4] dines@idiap.ch
[*] *Centre for Speech Technology Research,*
University of Edinburgh, U.K.
[2] jyamagis@inf.ed.ac.uk

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. By contrast, with just a single parameter, VTLN captures very few speaker specific characteristics when compared to the available linear transform based adaptation techniques. This paper proposes that the merits of VTLN can be combined with those of linear transform based adaptation technique in a Bayesian framework, where VTLN is used as the prior information. A novel technique of 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

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 previously developed for automatic speech recognition (ASR) [2]. The maximum likelihood linear transformation (MLLT) based adaptation techniques entail linear transformation of the means and variances of an HMM to match the characteristics of the speech for a given speaker. 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 in order to estimate 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], [5]. It is inspired from the physical observation that the vocal tract length (VTL) varies across different speakers in the range of around 18 cm in males to around 13 cm in females. The formant frequency positions are inversely proportional to VTL, and hence can vary around 25% [6]. 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. [7] using a grid search. An efficient implementation of VTLN using expectation maximization (EM) with Brent search optimization is presented by Saheer et.al. [3].

VTLN has been previously combined with constrained MLLR (CMLLR) for rapid adaptation in ASR [8]. The linear VTLN used by Breslin et.al. [8] is based on a set of precomputed quantised transforms unlike the gradient descent approach used in this paper. Also, a count smoothing framework is used to incorporate the prior information since a conjugate prior distribution is not possible for CMLLR. The incorporation of prior information can be achieved in

a more subtle and efficient way by using constrained structural maximum a posteriori linear regression (CSMAPLR). CSMAPLR itself is a more robust adaptation framework when compared to CMLLR, and inclusion of an efficiently estimated VTLN prior matrix for the root node can further enhance the performance. The structural framework helps propagate the prior information through the various levels of the regression tree in an effective way.

CSMAPLR has been proposed as a robust speaker adaptation technique for statistical parametric speech synthesis [2]. The maximum a posteriori (MAP) criterion can better estimate linear transforms when less adaptation data is available. The structural MAP (SMAP) technique uses an identity matrix as the initial prior distribution (at the root node) and uses a tree structure to propagate this prior to different classes of transforms. 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 little data is available. This paper presents the details of a novel technique of using VTLN as a prior in the Bayesian adaptation framework for CSMAPLR.

II. Theory

A. 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) [9], which makes use of a bilinear transform to achieve a frequency warp[1]. Since MGCEP already includes a bilinear transform as its spectral warping function to approximate the mel auditory scale, a bilinear transform-based VTLN proposed by Pitz and Ney [10] 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

[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.

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 [10]:

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

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

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

We may represent the linear transformation in the vector form

$$\boldsymbol{x}_\alpha = \boldsymbol{A}_\alpha \boldsymbol{x} \tag{9}$$

where $\boldsymbol{x}_\alpha = (\tilde{c}_1, \cdots, \tilde{c}_M)^\top$ and $\boldsymbol{x} = (c_1, \cdots, c_K)^\top$. The transform may also be directly applied to the dynamic features of the cepstra as well; 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 [6]:

$$\widehat{\alpha}_s = \underset{\alpha}{\arg\max} \, P(\boldsymbol{x}_{\alpha_s} \mid \Theta, \alpha_s, w_s) \tag{10}$$

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 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 [3], [11], which transforms the spectral features as follows

$$\tilde{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} = \boldsymbol{W}\boldsymbol{\xi}. \tag{11}$$

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.

B. CSMAPLR

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

$$\widehat{\boldsymbol{W}}_s = \underset{\boldsymbol{W}}{\operatorname{argmax}} \, P(\boldsymbol{x} \mid \Theta, \boldsymbol{W}_s, w_s) \, P(\boldsymbol{W}_s) \qquad (12)$$

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 $P(\boldsymbol{W})$:

$$P(\boldsymbol{W}) \propto |\boldsymbol{\Omega}|^{-\frac{L+1}{2}} |\boldsymbol{\Psi}|^{-\frac{L}{2}}$$
$$\exp\left[-\frac{1}{2}\mathrm{tr}(\boldsymbol{W}-\boldsymbol{H})^{\top}\boldsymbol{\Omega}^{-1}(\boldsymbol{W}-\boldsymbol{H})\boldsymbol{\Psi}^{-1}\right] \qquad (13)$$

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 scaling.

C. Using VTLN as CSMAPLR Prior

The VTLN transformation presented in this paper can be considered as a very constrained form of CMLLR/CSMAPLR, only requiring estimation of a single parameter representing the VTL of the speaker. This parameter normally gives some measure of the vocal tract length, 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 (refer Figure 1). Hence, for the prior distribution at the root node, we may set the hyperparameter $\boldsymbol{H}$ as

$$\boldsymbol{H}_{\mathrm{VTLN}} = [\boldsymbol{A}_\alpha, \boldsymbol{0}] \qquad (14)$$

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
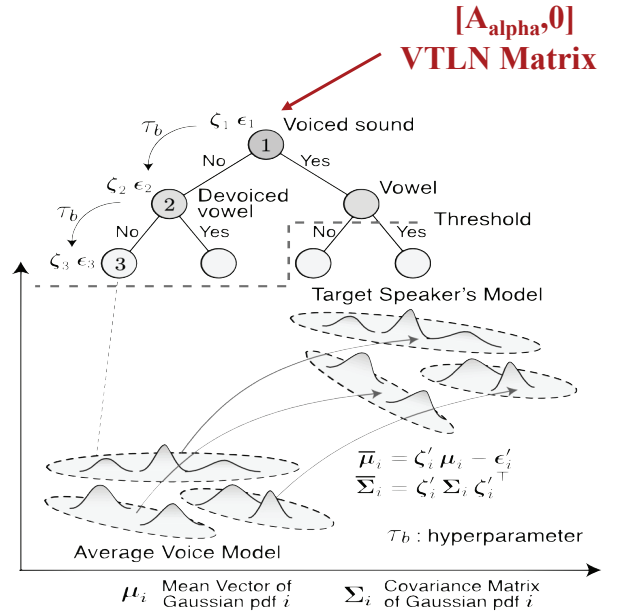

Fig. 1: Identity matrix is replaced with VTLN transformation matrix as prior for the root node of the CSMAPLR transformation.

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.

The gender 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.

III. EVALUATIONS WITH VTLN AS PRIOR

A. 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, $\log F_0$, five-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 UK English speech corpus including 5 hours of clean speech data uttered by an RP professional narrator. The evaluation experiments are performed on another UK English test speaker. Subjective listening tests are performed using
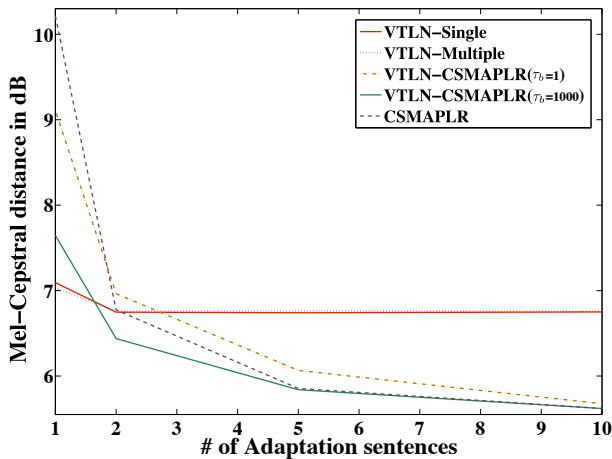
Fig. 2: Mel-cepstral distances of VTLN, CSMAPLR, and the proposed VTLN-CSMAPLR techniques.

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 number of subjects was 40.

The subjective tests are based on mean opinion scores (MOS) of naturalness and speaker similarity. The synthesized utterances are rated on a 5-point scale, 5 being "completely natural" or "sounds exactly like the target speaker" and 1 being "completely unnatural" or "sounds like a totally different speaker". 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 of the values derived from the natural speech, and can be viewed as an approximation to the log spectral distortion measure according to Parserval's theorem.

B. Results and Discussion

One hundred sentences were synthesized for objective evaluations for the test speaker. The values of the MCD scores for different amounts of adaptation data are plotted in the Figure 2. The figure shows the MCD score for two different scale factors, 1 and 1000. The figure also compares the objective scores with the VTLN adaptation technique based on a single warping factor or regression class based multiple warping factors. The objective results first 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 even 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 10dB MCD with only one adaptation utterance. Finally, the objective results clearly show that the proposed VTLN-CSMAPLR technique with scale factor 1000 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 and its distortion is around 6.5dB.

The listening tests were performed with 1, 2 and 5 adaptation sentences. The evaluation results of the listening tests are shown in Figure 3. From the speaker similarity results, we can see that VTLN works best when the number of adaptation sentences is one or two and also that VTLN-CSMAPLR outperforms CSMAPLR with one or two adaptation sentences. There is no significant difference among the adaptation methods with five adaptation sentences. 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 or two adaptation sentences. From the intelligibility evaluation, we observe that there is no significant difference between VTLN and VTLN-CSMAPLR with two and five sentences, but, on the other hand, we can see that CSMAPLR has significantly degraded intelligibility with one adaptation sentence. Although there are minor inconsistencies with the MCD results, we conclude that the VTLN prior can significantly improve the CSMAPLR adaptation performance when the adaptation data is very limited.

IV. CONCLUSIONS

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.

ACKNOWLEDGMENT

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.
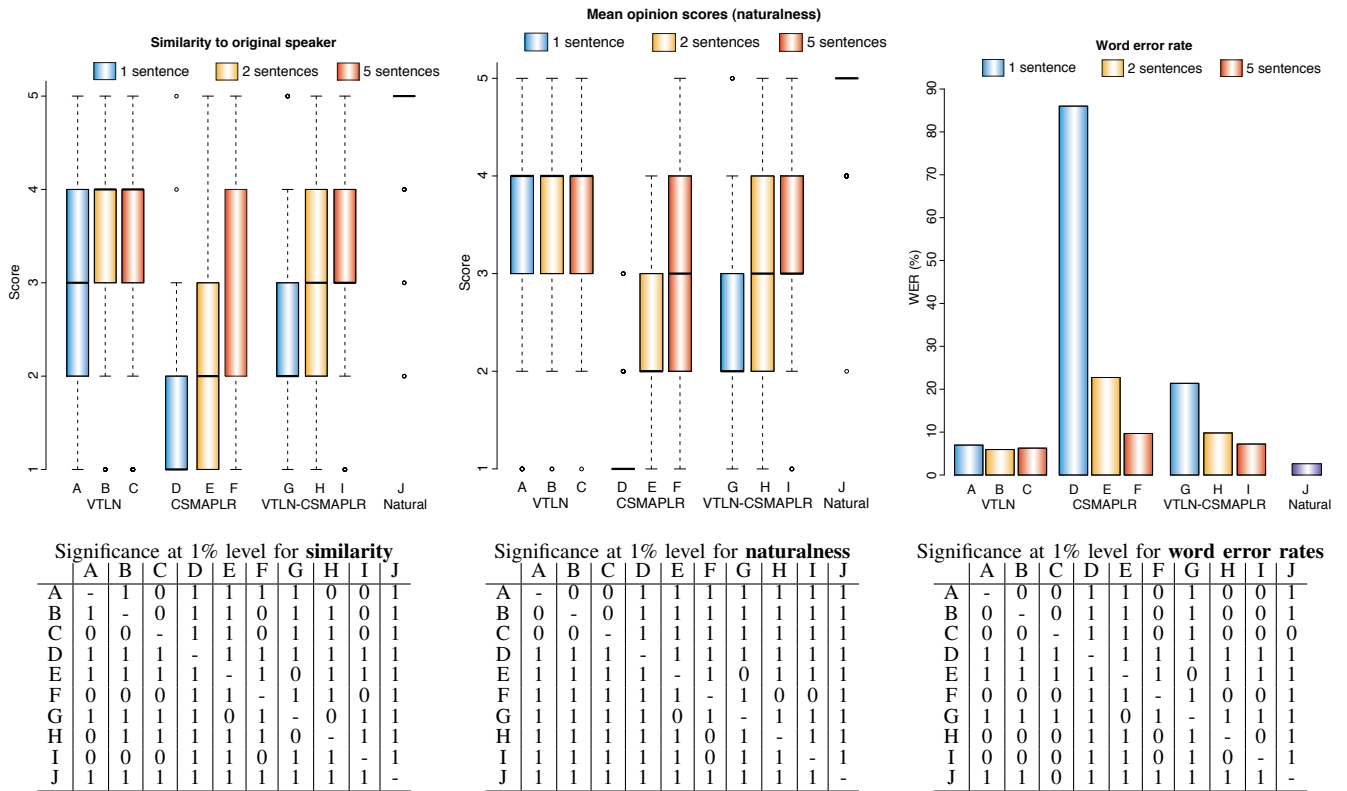
Fig. 3: Listening tests results. There are three columns of plots and tables which are, from left to right, similarity to original speaker, mean opinion score for naturalness, and intelligibility. The similarity and naturalness plots on the upper row are box plots where the median is represented by a solid bar across a box showing the quartiles and whiskers extend to 1.5 times the inter-quartile range. The 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.

Significance at 1% level for **similarity**

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

Significance at 1% level for **naturalness**

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

Significance at 1% level for **word error rates**

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

[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, Sep. 2010, pp. 224–229.

[4] 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.

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

[6] 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.

[7] 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.

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

[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] 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.

[12] 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.