A FREQUENCY-WEIGHTED POST-FILTERING TRANSFORM FOR COMPENSATION OF THE OVER-SMOOTHING EFFECT IN HMM-BASED SPEECH SYNTHESIS

Florian Eyben[1]*, Yannis Agiomyrgiannakis*

Google

eyben@tum.de, agios@google.com

ABSTRACT

Over-smoothing is one of the major sources of quality degradation in statistical parametric speech synthesis. Many methods have been proposed to compensate over-smoothing with the speech parameter generation algorithm considering Global Variance (GV) being one of the most successfull. This paper models over-smoothing as a radial relocation of poles and zeros of the spectral envelope towards the origin of the z-plane and uses radial scaling to enhance spectral peaks and to deepen spectral valeys. The radial scaling technique is improved by introducing over-emphasis, spectral-tilt compensation and frequency weighting. Listening test results indicate that the proposed method is 11%-13% more preferable than GV while it has less algorithmic delay (only 5 ms) and computational complexity.

Index Terms— speech synthesis, hidden Markov model, post-filtering, mel-frequency cepstrum, global variance, radial scaling

1. INTRODUCTION

Statistical Parametric Speech Synthesis [1] (SPSS) relies on a mapping between linguistic features derived from text and acoustic features that parameterize the speech signal. The mapping itself has an over-smoothing effect to the generated acoustic features which can be attributed to the fact that linguistic features are not descriptive enough to discriminate between different realizations of the acoustic vectors. In SPSS based on Hidden Markov Models (HMMs), the mapping between the linguistic and acoustic features is made via decision trees. The decision tree maps the linguistic features (also referred to as *linguistic contexts*) to a state of the context-dependent HMM that describes the distribution of the acoustic features. Alternatively, one may use a neural network to obtain a distribution of the acoustic features [2], [3] or restricted Boltzman machines [4]. Speech synthesis is made by mapping the input text to the linguistic features and the linguistic features to the acoustic Gaussian distributions. A sequence of Gaussian distributions is obtained, typically, one per 5 ms of speech. Then, a trajectory of acoustic features is derived from the sequence of acouastic models, a process that is referred to as parameter generation [5]. The trajectory is effectively smoothed by considering delta- and delta-delta- acousting features during parameter generation [1]. Overall, statistical mapping and parameter generation have an over-smoothing effect to the generated spectral sequence [6].

A number of methods have been proposed to reduce the over-smoothing effect. Post-filtering methods originating from speech coding have proven to be quite effective [7], [8], [9]. Post-filtering increases the peak-to-valey range of the formants in a way that is adapted to the generated spectral envelope but ignores the characteristics of the trajectory of the generated parameters. Toda et al. [6] observed that the variance of the generated parameters over time is less than the variance of the original parameters and incorporated a compensating term to the parameter generation phase. This method and its offsprings are referred to as Global Variance (GV) [10]. The method was originally made for Mel Cepstra (MCEP) but it can be adapted to LSPs (Line Spectral Pairs) as well [11], [12]. Global variance can be seen as a form of variance adaptation along the trajectory. An extension of this idea is to use histogram equalization techniques [13]. Since post-filtering increases the dynamic range of the spectral parameters along the frequency axis while GV increases the dynamic range of the spectral envelopes along the time-axis, it is possible to have a configuration that uses both methods advantageously.

In z-domain, over-smoothing effectively moves spectral envelope poles/zeros inwards and away from the unit-circle. Radial scaling of poles and zeros has been used extensively in speech coding for the construction of perceptual weighting filters in Code-Excited Linear Prediction (CELP) coders [7], [14] and regularization [15]. The transform moves poles and zeros along the radius of the unit-circle by multiplying their radii with a factor ρ. Radial scaling corresponds to the z-domain transform $z' = \rho/z$. In [16], Sorin et al. proposed to use radial scaling for the enhancement of spectral envelopes in multi-form HMM synthesis and estimates ρ from pairs of measured and synthesized spectra so that 2-nd order moments of synthesized spectra are close to 2-nd moments of measured spectra in log-domain. Scale factors are considered to be context-dependent, thus a different ρ is computed for every HMM state.

The latter method has the advantage of providing a single-parameter transform for the enhancement of spectral envelopes and in our experiments it did provide improvement over the baseline but could not rival GV. A drawback of this method is that it does not consider spectral-tilt. Spectral tilt removal prior formant enhancement is a common step in speech coding [17], [14]. Another drawback is that it moves poles and zeros equally towards the unit-circle assuming that over-smoothing is uniformly distributed in frequency. However, higher frequencies seem to be more smeared than lower frequencies, which can be partially attributed to mel-scale frequency warping in the spectral envelope parameterization.

This paper presents a novel radial scaling method for over-smoothing compensation in HMM-based SPSS. The proposed method uses context-dependent radial scaling [16] and introduces three new ideas: over-emphasis, spectral-tilt compensation and frequency weighting. Over-emphasis allows us to extend the amount of emphasis beyond the one estimated, spectral-tilt compensation flattens the spectrum prior post-filtering and frequency weighting allows us to avoid applying over-emphasis in lower frequencies to the benefit of over-smeared higher frequencies.

[1]The first author was an intern at Google London (UK) at the time the experiments were conducted.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

The remainder of this paper is structured as follows: Section 2 describes the effect of radial scaling in Auto-Regressive Moving-Average (ARMA) filters. Section 3 discusses spectral averaging from the perspective of the source-filter theory of speech production. Section 4 describes the proposed method and Section 5 explains how to obtain the spectral-tilt vectors and frequency weighting matrices used in Section 4. Listening test results are presented in Section 6 and a conclusion is given in Section 7.

2. RADIAL SCALING IN ARMA FILTERS

The spectral envelope of speech is conveniently modelled as a minimum phase Auto-regressive Moving Average (ARMA) filter $H(z)$ with the following transfer function in z-domain:

$$H(z) = A\frac{\prod_{m=1}^{M}\left(1-z^{-1}z_m\right)}{\prod_{k=1}^{K}\left(1-z^{-1}p_k\right)}, |p_k|<1, |z_m|<1, \quad (1)$$

where A is a gain factor and p_k and z_m are the poles and zeros, respectively. The transform $z' = \rho/z$ radially moves p_k and z_m along the unit circle. The transformed transfer function has zeros $z'_m = \rho z_m$ and poles $p'_k = \rho p_k$. When $\rho > 1.0$ the movement is made towards the unit circle, increasing spectral peaks and decreasing spectral valleys. For speech signals, this corresponds to a reduction of format bandwidth causing formants to become sharper and valleys deeper. When $\rho < 1.0$ the movement is made towards the origin with a smearing effect to the spectrum. In speech coding, this transform is referred to as *bandwidth expansion* because it increases formant bandwidth [15], [7]. Taking the complex logarithm of $H(z)$ gives:

$$\log(H(z)) = \log(A) + \sum_{m=1}^{M}\log\left(1-z^{-1}z_m\right) - \sum_{k=1}^{K}\log\left(1-z^{-1}p_k\right), \quad (2)$$

and using the series expansion:

$$\log(1-\alpha z^{-1}) = -\sum_{n=1}^{\infty}\frac{\alpha^n}{n}z^{-n}, |\alpha z^{-1}|<1, \quad (3)$$

the (real) cepstrum is:

$$c[n] = \log(A)\delta[n] + \sum_{k=1}^{K}\frac{p_k^n}{n}u[n] - \sum_{m=1}^{M}\frac{z_m^n}{n}u[n], \quad (4)$$

where $u[n]$ is the step function ($u[n]=1$ when $n>0$ and zero otherwise) and $\delta[n]$ is the delta function ($\delta[n]=1$ when $n=0$ and zero otherwise).

If we now apply a common linear scale factor ρ to all poles and all zeros, we get:

$$\rho^n c[n] = \log(A)\delta[n] + \sum_{k=1}^{K}\frac{\rho^n p_k^n}{n}u[n] - \sum_{m=1}^{M}\frac{\rho^n z_m^n}{n}u[n], \quad (5)$$

which provides a convenient way to perform radial scaling with a constant scale factor directly in the cepstral domain:

$$c'[n] = \rho^n c[n]. \quad (6)$$

It is simple to show that radial scaling also corresponds to exponential weighting of the auto-regressive and moving average coefficients and to exponential windowing of the impulse response [15]. Care, however, has to be taken so that radial scaling does not render the filter unstable; having all poles and zeros within the unit-circle requires $|\rho^k p_k|<1$, $|\rho^m z_m|<1$. Further, note that the case where the frequency axis is warped (i.e., along Mel-Scale) does not effect the nature of the effect as pole/zero movement is made along the radius [16]. Interestingly enough, neither we or [16] identified a need for energy normalization during radial scaling. This deserves further investigation, though.

3. SPECTRAL AVERAGING AND THE SOURCE-FILTER MODEL OF SPEECH PRODUCTION

According to the source-filter model of speech production, the speech signal $s[n]$ can be described as a LTI (linear time invariant) system with the following transfer function in z-domain [17]:

$$S(z) = E(z)G(z)V(z)L(z) \quad (7)$$

where $E(z)$ models the excitation of the system (i.e. a pulse train), $L(z)$ models the radiating effect of the lips, $G(z)$ models the glottal source and $V(z)$ the vocal tract. Modern vocoders model the spectral envelope using minimum-phase assumptions, effectively ignoring the anti-causal behavior of the glottal source during the parameter extraction phase. Thus, the spectral envelope of the speech signal is modelled by $H(z) = G(z)V(z)L(z)$. The complex logarithm of the spectral envelope is:

$$\log(H(z)) = \log(G(z)) + \log(V(z)) + \log(L(z)). \quad (8)$$

The averaging effect that results from the mapping between the linguistic features $\boldsymbol{f}$ and the spectral envelopes can be modeled as the conditional expectation of $\log(|H(e^{j\omega})|)$ given the linguistic features $\boldsymbol{f}$:

$$E\left\{\log(|H(e^{j\omega})|)|\boldsymbol{f}\right\} = E\left\{\log(|G(e^{j\omega})|)|\boldsymbol{f}\right\} + E\left\{\log(|V(e^{j\omega})|)|\boldsymbol{f}\right\} + E\left\{\log(|L(e^{j\omega})|)|\boldsymbol{f}\right\}. \quad (9)$$

Now, it is important to note is that the components of the spectral envelope are not equally effected by the averaging. The radiation effect from the lips is relatively constant and the glottal source is related to changes in phonation and voice quality, which are kept relatively constant during recordings while the variation of the vocal tract can be much higher because it is inflicted by the lack of information in linguistic features. Professional voice talents used in modern speech database recordings are able to speak with a constant speaking style and voice quality during the recordings. Further investigation is, however, needed to determine the level of variation of each component within a linguistic context, although some relationship with pitch is already well documented [18].

4. CONTEXT-DEPENDENT RADIAL SCALING WITH FREQUENCY WEIGHTING

The enhancement method we propose is applicable to HMM-based Text to Speech (TTS) systems where the spectral envelope of the synthesized speech signal is parameterized using MCEP. The method uses radial scaling [16] to increase the peak-to-valley spectral distance around formants. Figure 1 shows an example of a spectral envelope with three distinct peaks and three different scale factors $\rho \geq 1$. As the scale factor increases, the peaks become higher and sharper (poles move towards unit-circle) and the valleys

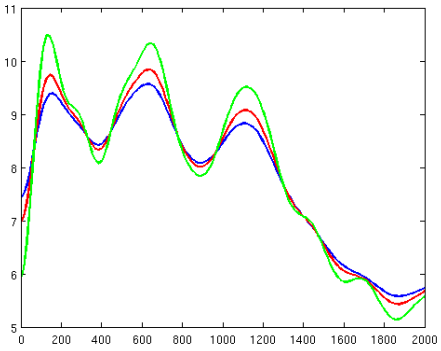

Fig. 1. Spectral envelope of phoneme */aa/* with three distinct peaks. Effect of scale factor: blue line (bottom line), unfiltered; red line (middle) $\rho = 1.0163$; green line (top): $\rho = 1.0229$.

become lower and narrower (zeros move towards unit-circle). Radial scaling is seen as a compensation transform for over-smoothing. A scale factor ρ is individually estimated for each state of the HMM. Each HMM state is modelled using a multivariate Gaussian distribution with a diagonal covariance matrix. The estimation is made as follows:

An HMM-based TTS synthesizer is trained. The training dataset utterances are synthesized with that HMM-based synthesiser. The generated spectral envelopes (MCEP) are stored prior vocoding. Let $\boldsymbol{c}_{data} = [c_{data}[1], ..., c_{data}[P]]^T$ and $\boldsymbol{c}_{gen} = [c_{gen}[1], ..., c_{gen}[P]]^T$ be the training-set MCEP vectors and the generated-set MCEP vectors, respectively. The generated-set MCEP sequences are aligned to the training-set MCEP sequences and state-level correspondances are extracted. For each HMM state, the corresponding *empirical 2-nd moments* for the p-dimension are computed as $M_{data}[p]^2 = E\{c_{data}[p]^2\}$ and $M_{gen}[p]^2 = E\{c_{gen}[p]^2\}, p = 1, ..., P$ from the training data vectors and the generated data vectors, respectively. The computation of the moments is made using only the vectors that belong to the corresponding HMM state. Each moment vector is smoothed using a 5-tap moving average filter along the vector dimensions $1, ..., P$. The scale factor of the each HMM state is computed using the following formulas [16]:

$$r_p = \sqrt{\frac{M_{data}[p]^2}{M_{gen}[p]^2}}, p = 1, ..., P \tag{10}$$

$$\log(\rho) = \frac{\sum_{p=1}^{P} p \log(r_p)}{\sum_{p=1}^{P} p^2}, p = 1, ..., P \tag{11}$$

where $r_p, p = 1, ..., P$ is the L_2-norm ratio sequence. The formula is empirically derived from observations of L_2-norm ratio sequences and corresponds to a linear regression fit in the logarithmic sequence $\log(r_p), p = 1, ..., P$. The solution can also be interpreted as a least-squares fit of the moments in the log-domain.

An example of radial scaling applied to the phoneme */aa/* is given in Figure 1. The figure depicts the original spectral envelope as produced by the HMM synthesiser and the version enhanced by the automatically estimated scale factor ρ. The enhancement emphasizes peaks and valleys, however, the peak-to-valley distance is still smaller than that of the human speech samples. Speech synthesized using Equation (11) sounds improved over the baseline, but not as good as the one obtained via Global Variance. In order to further improve quality, we propose to apply an over-emphasis factor ζ to the scale factors ρ:

$$\hat{\rho} = \rho^{\zeta}. \tag{12}$$

The result of the over-emphasis is also shown in Figure 1 (green line; line with most variation). We experimentally determined that good values for the over-emphasis factor ζ are in $[1.0, 1.8]$.

The figure demonstrates that radial scaling affects the left-most peak around the fundamental frequency, which is the result of the interplay between the glottal formant and the first format. If radial scaling is applied uniformly to all frequencies, the latter peak will also be emphasized leading to two sorts of artifacts: first, the synthesis filter may become unstable and second, the glottal formant is emphasized which alters voice quality. The first artifact practically limits the amount of over-emphasis that can be applied. The second artifact reduces the *naturalness* of speech as the glottal formant becomes disproportionally dominant. On the other hand, we observed that over-emphasizing higher frequencies is beneficial because it makes speech sound more "present" and "clear"; terms that are hereby used in all caution as there is no standardized or dominant terminology for the description of sound quality. From the signal processing point of view, it is reasonable to expect that higher formants are more smeared due to Mel-scale frequency warping in MCEP.

We propose to address these issues using tilt-removal and frequency weighting in the radial scaling. Tilt-removal is a typical step that is made prior to the enhancement step in post-filtering. A post-filter is essentially a normalized, compressed version of the spectrally flattened vocal tract spectral envelope ([17], ch 12). In this work, we used a simple fixed deemphasis filter:

$$D(z) = (1 + \gamma)^{-1} \left(1 - \gamma z^{-1}\right), \tag{13}$$

with $\gamma = 0.96$ for 16 kHz sampling rate. Let $\boldsymbol{\delta}$ be an approximation of this filter in the MCEP domain and $\boldsymbol{c}$ be the MCEP spectral envelope that we want to enhance. First we remove the spectral tilt from $\boldsymbol{c}$:

$$\boldsymbol{c}' = \boldsymbol{c} + \boldsymbol{\delta}. \tag{14}$$

Then we apply a radial scaling transform as a simple matrix-vector multiplication:

$$\boldsymbol{c}'' = \boldsymbol{W}\boldsymbol{c}', \tag{15}$$

where $\boldsymbol{W} = \text{diag}\left\{[\rho^{\zeta}, \rho^{2\zeta}, ..., \rho^{P\zeta}]\right\}$. By defining $\boldsymbol{F} = \boldsymbol{W} - \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix, we can express $\boldsymbol{c}''$ as

$$\boldsymbol{c}'' = \boldsymbol{W}\boldsymbol{c}' = (\boldsymbol{I} + \boldsymbol{F})\boldsymbol{c}' = \boldsymbol{c}' + \boldsymbol{F}\boldsymbol{c}'. \tag{16}$$

The above equation expresses the radial scaling process as a filtering of the de-tilted spectral envelope $\boldsymbol{c}'$ with a post-filter $\boldsymbol{e} = \boldsymbol{F}\boldsymbol{c}'$. However, the post-filter is still applied uniformly to all frequencies. Emphasizing lower frequencies can be avoided by applying frequency weighting to the post-filter term $\boldsymbol{e}$. In the MCEP domain, this can be done via a matrix-vector multiplication with a frequency weighting matrix $\boldsymbol{B}$:

$$\boldsymbol{e}' = \boldsymbol{B}\boldsymbol{e} = \boldsymbol{B}\boldsymbol{F}\boldsymbol{c}'. \tag{17}$$

Details regarding the construction of matrix $\boldsymbol{B}$ and vector $\boldsymbol{\delta}$ are separately presented in Section 5 for clarity. The enhanced spectral envelope is obtained by combining the equations above and re-introducing the spectral tilt after the frequency-weighted radial scaling:

$$\hat{\boldsymbol{c}} = \boldsymbol{c} + \boldsymbol{B}\boldsymbol{F}(\boldsymbol{c} + \boldsymbol{\delta}). \tag{18}$$

The implementation goes as follows: during training we estimate and store a scale factor for each Gaussian state of the

HMM; during synthesis, we retrieve a scale factor for each 5ms frame of speech, apply over-emphasis (eq. 12) and smooth the sequence of emphasized scale factors with a zero-phase 3-tap filter: $[0.15, 0.70, 0.15]$. Then, we use this factor to perform the post-filtering. Thus, the proposed method has an algorithmic delay of 5 ms (1 frame) which is much smaller than the algorithmic delay of GV which requires the whole utterance [19]. The complexity of the proposed method is also less than GVs' one.

5. SPECTRAL TILT FILTER AND FREQUENCY WEIGHTING MATRIX IN MCEP DOMAIN

This section shows how to construct the deemphasis filter $\boldsymbol{\delta}$ and the frequency weighting matrix $\boldsymbol{B}$ in MCEP domain. Let $\boldsymbol{c} \in R^P$ be a MCEP vector and $\boldsymbol{c}_{\log} \in R^N$ be the corresponding log-amplitude spectral envelope that is sampled at the DCT-II (Discrete Cosine Transformation) frequency grid $\omega_n = (\pi/N) * (n - 0.5), n = 1, ..., N$. The log-amplitude spectrum can be obtained from MCEP parameterization using [20]:

$$c_{\log}[n] = \sum_{p=1}^{P} c[p] \cos\left(\tilde{\omega_n}(p-1)\right), \tag{19}$$

where $c_{\log}[n]$ and $c[p]$ are the n-th and p-th element of the corresponding vectors, while

$$\tilde{\omega}_n = \tan^{-1} \frac{(1-\alpha^2)\sin(\omega_n)}{(1+\alpha^2)\cos(\omega_n) - 2\alpha}. \tag{20}$$

For a fixed frequency grid (i. e., the DCT-II) this can be written in matrix form:

$$\boldsymbol{c}_{\log} = \boldsymbol{C}_{lm}\boldsymbol{c}, \tag{21}$$

where $\boldsymbol{C}_{lm}$ is an N-by-P MCEP-to-log-amplitude conversion matrix with elements $C_{lm}[n,p] = \cos\left(\tilde{\omega_n}(p-1)\right), n = \{1, ..., N\}$, $p = \{1, ..., P\}$. The reverse operation (computing the MCEP parameters from a log-amplitude frequency response sampled at the DCT-II frequency grid) can be made using the pseudo-inverse matrix $\boldsymbol{C}_{ml} = \mathrm{pinv}(\boldsymbol{C}_{lm})$. Matrix $\boldsymbol{C}_{lm}$ is a frame expansion that goes from P parameters (i. e., 40) to $N \gg P$ parameters (i. e., 1024), thus, the pseudo-inverse is optimal in the Mean Squared Error (MSE) sense because it projects the log-amplitude spectrum back to the subspace of the MCEP.

Let $\boldsymbol{\delta}_{\log}$ be the log-amplitude frequency response of $D(z)$ (eq. 13) at frequencies $\omega_n, n = 1, ..., N$. The deemphasis filter then in MCEP is $\boldsymbol{\delta} = \boldsymbol{C}_{ml}\boldsymbol{\delta}_{\log}$.

The frequency weighting matrix $\boldsymbol{B}$ is constructed as:

$$\boldsymbol{B} = \boldsymbol{C}_{ml}\boldsymbol{\Delta}\boldsymbol{C}_{lm}, \tag{22}$$

where $\boldsymbol{\Delta}$ is a diagonal matrix with elements:

$$\Delta[n,n] = \begin{cases} \frac{\omega_n}{\omega_c}, & \text{if } \omega_n \leq \omega_c \\ 1, & \text{otherwise.} \end{cases} \tag{23}$$

Thereby ω_c is the transition threshold placed at 500 Hz.

6. EVALUATION

We conducted crowd-sourced listening tests to evaluate subjectively the proposed system. The experimental conditions for the HMM systems were similar to those presented in [2], Section 4.1. The tests were conducted as A/B preference tests, where the raters were presented two samples of speech (each the same sentence) and had to state their preference on a 5-point scale. The samples were presented in random order to avoid ordering bias. One male and one female British English voice were evaluated using 30 synthesized utterances. Each pair was evaluated by roughly the same number of raters. The tests were run on Amazon's Mechanical Turk platform. Each rater was presented to a random subset of 8 A-B tests in one session. Since we cannot enforce raters to use headphones, it is best to bias them and let them declare their preferences. Further, having both speaker and headphone usage in the subjective evaluation is more representative of real world usage.

Table 1. *Subjective preference (Pr.) test results. Comparison of baseline system '-' (no post processing of parameters) with baseline + global variance (gv), baseline + proposed post-filter approach with over-emphasis factor 1.4 and de-emphasis transform (pf), and baseline + global variance and proposed post-filter (gv + pf). Columns A and B: type of systems compared in paired listening test; Pr. A/B: percentage of subjects preferring system A over B (Pr. A) or B over A (Pr. B). Last two rows: comparison of listener preference when listening to samples via speakers (S) or headphones (H). Last column contains the number of ratings per test N_s, N_h that is made with speakers and headphones, respectively.*

Voice	A	B	A [%]	B [%]	p-val.(B)	N_h	N_s
F	gv	pf	6.8	**65.8**	< 0.001	256	104
F	-	pf	10.8	**23.3**	< 0.001	152	88
M	gv	pf	24.2	**35.1**	< 0.002	376	240
M	-	pf	15.8	**48.3**	< 0.001	160	72
M	gv	gv + pf	**29.2**	18.3	> 0.991	152	88
M (S)	-	pf	11.1	**23.6**	< 0.07	-	-
M (H)	-	pf	18.8	**61.9**	< 0.001	-	-

The results of the preference tests are given in Table 1. Thereby the speech synthesized by the same system with and without the proposed enhancement method was compared to a baseline Mel-Cepstral HMM system with no post enhancement of the parameters and global variance maximation based enhancement of parameters. We can observe that the proposed post-filtering method has consistently higher preference over GV. The preference is significantly higher for the female speaker because the corresponding speech corpus exhibits high variability that effects the performanceof GV. When GV is de-activated and simple post-filtering is used, the preference falls to 13%. For the male speaker, the preference is 11%. All preference values are statistically significant. We found that the cascaded operation of applying our post-filtering to GV-derived spectral envelopes degraded quality over GV. In that experiment, the scale factors were computed using GV-derived spectral envelopes. Qualitatively, the formants were over-emphasized.

7. CONCLUSION

We propose a fast, lightweight, context-dependent post-filtering method for improvement of speech quality in HMM-based TTS systems. The method improves radial scaling techniques by introducing spectral tilt removal prior modification and a novel frequency weighting mechanism. Listening tests indicate that the proposed method is preferred over the state-of-the-art method of Global Variance by 11% for a male voice and 13% over the baseline for a female voice for which GV does not work properly due to high variability present in the training data. The proposed method has only 5 ms algorithmic delay.

8. REFERENCES

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

[2] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. of ICASSP, Vancouver, Canada, May*, 2013, pp. 7962–7966.

[3] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "F0 contour prediction with a deep belief network-Gaussian process hybrid model," in *Proc. ICASSP*, 2013, pp. 6885–6889.

[4] Z.-H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted Boltzmann machines for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 2129–2139, 2013.

[5] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. of ICASSP*. IEEE, 2000, vol. 3, pp. 1315–1318.

[6] T. Tomoki and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Transactions on Information and Systems*, vol. 90, no. 5, pp. 816–824, 2007.

[7] K. Koishida, K. Tokuda, T. Kobayashi, and S. Imai, "CELP coding system based on mel-generalized cepstral analysis," in *Proc. of Spoken Language*. IEEE, 1996, vol. 1, pp. 318–321.

[8] Y. Takayoshi, *Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based Text-To-Speech systems*, Ph.D. thesis, Electrical and Computer Engineering, Nagoya Institute of Technology, 2002.

[9] Z.-H. Ling, Y.-J. Wu, Y.-P. Wang, L. Qin, and R.-H. Wang, "USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method," in *Blizzard Challenge Workshop*, 2006.

[10] M. Shannon and W. Byrne, "Fast, low-artifact speech synthesis considering global variance," *Proc. of ICASSP*, pp. 7869–7873, 2013.

[11] S. Pan, Y. Nankaku, K. Tokuda, and J. Tao, "Global variance modeling on frequency domain delta-LSP for HMM-based speech synthesis," in *Proc. of ICASSP*. IEEE, 2011, pp. 4716–4719.

[12] Z.-H. Ling, Y. Hu, and L.-R. Dai, "Global variance modeling on the log power spectrum of LSPs for HMM-based speech synthesis," in *Proc. of Interspeech*, 2010, pp. 825–828.

[13] Y. Ohtani, M. Tamura, M. Morita, T. Kagoshima, and M. Akamine, "Histogram-based spectral equalization for HMM-based speech synthesis using Mel-LSP," in *Proc. of Interspeech*, 2012, pp. 1155–1158.

[14] W. B. Kleijn and K. K. Paliwal, Eds., *Speech Coding and Synthesis*, Elsevier Science Inc., New York, NY, USA, 1995.

[15] P. Kabal, "Ill-conditioning and bandwidth expansion in linear prediction of speech," in *Proc. of ICASSP*. IEEE, 2003, vol. 1, pp. I–824.

[16] A. Sorin, S. Shechtman, and V. Pollet, "Uniform speech parameterization for multi-form segment synthesis," in *Proc. of ICASSP*, 2011, pp. 337–340.

[17] T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*, Pearson education, 2008.

[18] P. Keating and Y.-L. Shue, "Voice quality variation with fundamental frequency in english and mandarin.," *The Journal of the Acoustical Society of America*, vol. 126, no. 4, pp. 2221–2221, 2009.

[19] T. Toda and K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," in *Proc. of Interspeech*, 2005, pp. 2801–2804.

[20] K. Tokuda, T. Kobayashi, and S. Imai, "Recursive calculation of Mel-Cepstrum from LP coefficients," *Transactions of IEICE*, vol. 71, pp. 128–131, 1994.