Lombard Modified Text-to-Speech Synthesis for Improved Intelligibility: Submission for the Hurricane Challenge 2013

Antti Suni[1], *Reima Karhila*[2], *Tuomo Raitio*[2], *Mikko Kurimo*[2], *Martti Vainio*[1], *Paavo Alku*[2]

[1]Department of Behavioural Sciences, University of Helsinki, Helsinki, Finland
[2]Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
firstname.lastname@helsinki.fi, firstname.lastname@aalto.fi

Abstract

This paper describes modification of a TTS system for improving the intelligibility of speech in various noise conditions. First, the GlottHMM vocoder is used for training a voice with modal speech data. The vocoder and voice parameters are then modified to mimic the properties of Lombard effect based on a small amount of Lombard speech from the same speaker. More specifically, the durations are increased, fundamental frequency is raised, spectral tilt is decreased, the harmonic-to-noise ratio is increased, and a pressed glottal flow pulses are used in creating excitation. The formants of the speech are also enhanced and finally the speech is compressed in order to increase noise robustness of the voice. The evaluation results of the Hurricane Challenge 2013 indicate that the modified voice is mostly less intelligible than the unmodified natural speech, as expected, but more intelligible than the reference TTS voice, especially in the low SNR conditions.

Index Terms: Hurricane challenge, speech synthesis, GlottHMM, Lombard speech, intelligibility

1. Introduction

Due to recent advances in speech synthesis research, especially with hidden Markov model (HMM) based speech synthesis [1] and related hybrid approaches, the intelligibility of neutral synthesized speech in silent conditions is no longer an issue. Hence, the focus of research has been shifted towards more adverse listening conditions and noise robustness.

Within this line of research, two trends can be identified. Some studies attempt to maximize the intelligibility of synthetic speech by optimizing the speech parameters according to some objective measure such as perceptual evaluation of speech quality (PESQ) [2] or glimpse proportion (GP) [3]. Here, the modifications can become quite unintuitive compared to human speech. For example, in some noise conditions, lowering the fundamental frequency (F0) can be beneficial [4], which humans rarely do in the presence of noise. The underlying assumption seems to be that the human speech production apparatus has certain physical constraints and that some effects of increased vocal effort are unintended consequences which are not beneficial to intelligibility of speech.

On the other hand, some studies have focused on the modeling of human speech production in noise, or the Lombard effect, with the assumption that what humans do in the presence of noise, also increases the intelligibility of the modeled speech. The drawback of this approach is that either a priori knowledge of the relevant changes in acoustic features have to be utilized, such as in [5], or that fairly large labeled corpus of Lombard speech have to be acquired in order to use data driven techniques such as HMM-adaptation [6].

In this study, we attempted to model the Lombard effect by a simple voice conversion technique. Compared to the approach in [6], which uses HMM-based adaptation, the current method requires only few utterances of unlabeled Lombard speech data from the target speaker. Thus, the current method is closer to the method in [5], which uses the tuning of the vocoder parameters. However, a data-driven technique is used in this study, thus requiring less a priori information about the modifications and possibly yielding higher quality speech. The assumption is that the current method is able to generate very intelligible synthetic speech with less effort in voice building compared to previous methods.

2. Speech synthesis system

In this study, the GlottHMM statistical parametric speech synthesis system [7] is used. GlottHMM aims to accurately model the speech production mechanism by decomposing speech into the vocal tract filter and the voice source signal using glottal inverse filtering and emphasizing the modeling of the voice source. It is built on a basic framework of HMM-based speech synthesis system [8, 9], but it uses a distinct type of vocoder for parameterizing and synthesizing speech. GlottHMM has been shown to yield high-quality and intelligible synthetic speech [5, 7, 10, 11] and also highly intelligible Lombard modified speech [5] and Lombard speech [6]. Since the conception of GlottHMM [12], it has been constantly developed. It is most thoroughly described in [7], but further developments have been made to the system since then, of which some are described in [5, 10, 11, 13]. In order to give a concise description of the system used in this study, GlottHMM is shortly described next.

2.1. Parametrization

In the parameterization of speech with GlottHMM, speech is first high-pass filtered with a cut-off frequency of 70 Hz in order to remove possible low-frequency ripple. Then, speech signal is windowed and iterative adaptive inverse filtering (IAIF) [14, 15] is used to estimate the vocal tract filter and the voice source signal from speech. Linear prediction (LP) is used for spectral estimation in the IAIF method, and the estimated vocal tract filter is converted to line spectral frequencies (LSF) [16] for better representation of the LP information in HMM-training [17]. The estimated voice source signal is further parameterized with fundamental frequency (F0), harmonic-to-noise ratio (HNR) of five bands, and voice source LP spectrum, which is also converted to LSFs. Finally, the glottal closure instants (GCIs) of the voice source are detected and individual glottal flow pulses

are extracted, from which a library of pulses is constructed. The library consists of windowed two-period glottal flow derivative waveforms, which are linked with their corresponding voice source parameters. The analysis parameters of GlottHMM are shown in Table 1.

2.2. Synthesis

In synthesis, the extracted natural glottal flow pulses from the library can be utilized in several ways. Originally, only a single pulse was used by interpolating the pulse in time and scaling in amplitude [7, 12]. Alternatively, a pulse selection scheme [13] can be used to construct the voice source or the mean of the pulse library [18] can be used as a basis for synthesis. In this work, a pulse library scheme [13] is used for constructing the voice source. Glottal flow pulses from the library are selected for each time instant according to the target cost of the voice source parameters, ensuring the selection of pulses with appropriate voice source characteristic, and according to the concatenation cost of adjacent pulses, ensuring that the change between adjacent pulses is not too large, which could induce harsh voice quality. This process is optimized by the Viterbi search. After the selection, pulses are interpolated in time according to F0 and scaled in amplitude according to energy. Next, in order to control the degree of voicing of the excitation, noise is added to the pulse in five bands in spectral domain according the HNR measure. The pulses are then overlap-added and filtered with the vocal tract filter to create speech.

2.3. HMM training and parameter generation

A synthetic voice used in this work was trained from the provided modal speech training data: 3 hours (2861 sentences) of newspaper style sentences, modified rhyme test sentences, and Harvard sentences read by a male British English speaker. The speech data was downsampled to 16 kHz prior to feature extraction.

The voice was trained using methods and tools based on the EMIME 2010 Blizzard Entry [20], which in turn are based on the HTS speech synthesis toolkit [8, 9]. Context-dependent multi-space distribution hidden semi-Markov models (MSD-HSMM) were trained on the acoustic feature vectors described in Section 2.1. The stream structure was modified to accommodate the parameters of the GlottHMM vocoder. The vocal tract spectrum was trained together with energy in a single stream, while HNR and voice source spectrum were trained alone in separate streams, and F0 was modeled in multi-space distribution (MSD) stream. Model training was initialized from aligned data, and model clustering was done incrementally over several rounds. The full-context labels for training and test utterances were created with Festival system [21] using UniRPX phone set and rules.

Table 1: Speech features and the number of parameters.

Feature	Number of parameters
Vocal tract spectrum LSF	24
Energy	1
Fundamental frequency (F0)	1
Harmonic-to-noise ratio (HNR)	5
Voice source spectrum LSF	6
Pulse library	5275 pulses
+ corresponding parameters	5275×37

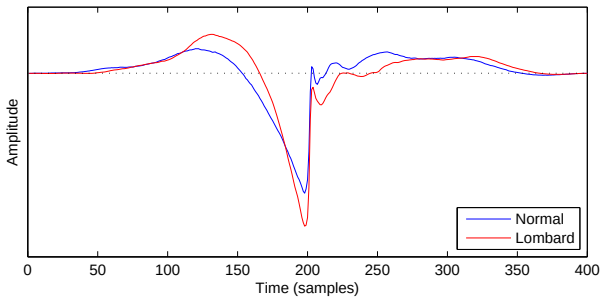

Figure 1: *Normalized average two-period glottal flow derivative waveforms of normal and Lombard speech.*

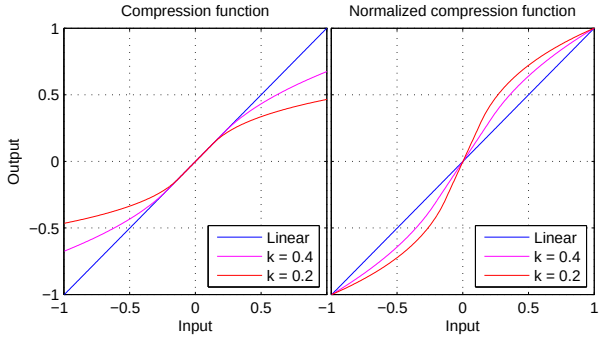

Figure 2: *Illustration of the compression function without (left) and with signal normalization (right).*

Since LSFs are highly correlated with each other, there are known problems with training of them. In order to avoid this problem, differential of the LSFs [19] were used both for vocal tract and voice source spectrum LSF parameterization. The differential LSF vector d_n is defined as

$$d_n = \begin{cases} l_n^{1/2} & \text{if } n = 0 \\ (l_n - l_{n-1})^{1/2} & \text{if } 0 < n < N \\ (\pi - l_{n-1})^{1/2} & \text{if } n = N \end{cases} \quad (1)$$

where l_n is the component n of the original LSF vector ($0 \leq n \leq N$). Vector d_n thus contains $N + 1$ values of which d_0 is the first LSF, d_n ($1 \leq n \leq N - 1$) are differences between the adjacent LSFs, and d_N is the distance from the last LSF to π. In order to get the distributions of the distances more Gaussian, the square root of the distances is taken. After the generation of the differential LSFs d'_n from HMMs, they are converted back to LSFs:

$$l'_n = \frac{d'^2_N}{\sum_{k=0}^{k=N-1} d'^2_k} \sum_{k=0}^{k=n} d'^2_k \quad (2)$$

where $0 \leq n \leq N - 1$. In this formula, the integrated differential-LSFs are equalized so that the distance from the last LSF to π equals to the original distance defined by d'^2_N.

3. Modifications for improved intelligibility

After the training of the voice with modal speech, an unsupervised method was used to transform the voice to a Lombard voice by modifying the synthesis parameters and tuning some

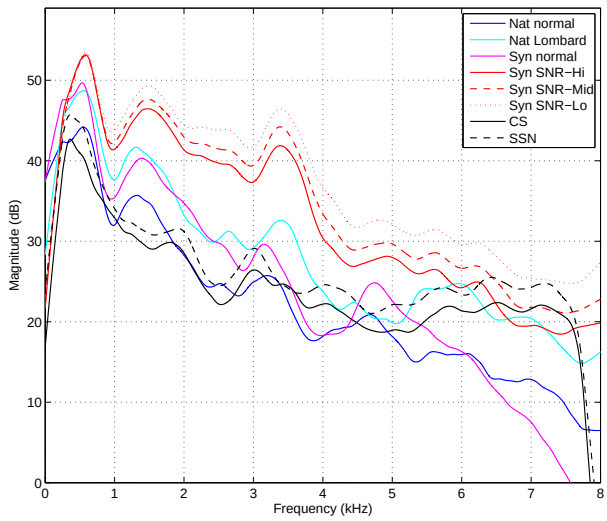

Figure 3: *Spectra of natural normal and Lombard speech, synthetic normal and three Lombard-modified voices designed for different SNRs, and competing speaker (CS) and speech-shaped noise (SSN) maskers.*

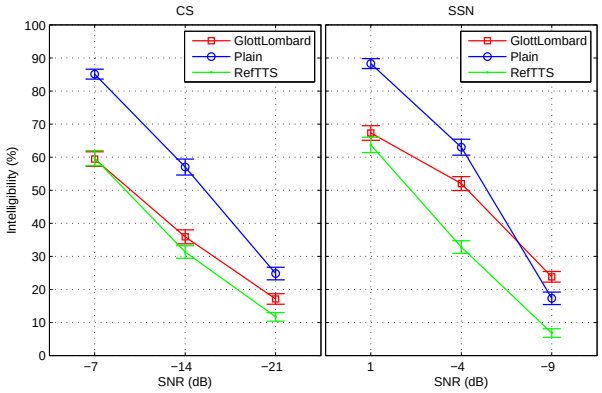

Figure 4: *Intelligibility of the proposed TTS voice (GlottLombard), original speech (Plain), and the reference TTS system (RefTTS) masked by competing speech (CS) and speech-shaped noise (SSN) with three different SNRs. Data is represented as means and standard errors over listeners.*

of the vocoder settings, thus requiring only a very small amount of unlabeled Lombard speech material for adaptation (20 utterances). The formant structure of speech was also enhanced in order to improve intelligibility. Phone durations were also increased by 10% to roughly match the corresponding unit lengthening observed in Lombard speech. Finally, speech signal was compressed to increase average loudness.

Neither noise type nor the maskers for individual utterances, provided by the organizers, were exploited in the modifications, as such detailed knowledge was considered somewhat artificial for the target domain of TTS.

3.1. Modification from normal voice to Lombard

A simple unsupervised voice conversion method was used to transform the plain synthetic voice to Lombard voice. First, a pulse library was built from 20 utterances of Lombard speech from the target speaker, consisting of 5275 glottal flow pulses along with the corresponding voice source (and also the vocal tract) parameter, described in Table 1. Figure 1 demonstrates the difference between the mean of the pulse libraries between normal and Lombard speech. In comparison to modal pulse, the Lombard pulse shows a shorter open phase and a more abrupt glottal closure.

In synthesis, the generated voice source parameters of each utterance were transformed to match the respective means and variances of Lombard speech, captured by the pulse library. This transformation was considered only for voiced speech. For each time instant, glottal flow pulses (of Lombard speech) were selected from the pulse library according to [13] (see Section 2.2), considering the *transformed* voice source features – F0, energy, voice source spectrum, and HNR – in target cost calculation. Concatenation cost was considered normally. Each individual glottal pulse was further modified in the spectral domain to emphasize high frequencies, thus decreasing the spectral tilt of the voice source even more.

The vocal tract spectrum was also transformed, but to a lesser degree, as the vowels were subject to distortion due to varying phonetic content of individual utterances. However, formant enhancement [22] was applied to a high degree to compensate for statistical smoothing as well as mimicking the clarity of Lombard speech.

The resulting speech signal was adaptively high-pass filtered in order to reduce the gain below the current F0 of speech. Finally, speech signals were compressed in order to increase average loudness using the following formula:

$$s'_n = \begin{cases} -(-s_n)^k + a & \text{if } s_n < -t \\ s_n^k - a & \text{if } s_n > t \end{cases} \quad (3)$$

where $t = (1/k)^{1/(k-1)}$ defines the starting point ($0 \leq t < 1$) of the compression, $a = t^k - t$ fits the compressed part of signal with the uncompressed one, and k is the compression coefficient ($0 \leq k < 1$). The smaller the value of k, the stronger the effect of compression. The idea of this compression is to increase the small signal values linearly and reduce the amplitude of the high peaks, which in effect will increase the overall gain of speech as the maximum value of the signal is normalized to 1. Figure 2 shows the compression function (left panel) with two different degrees of compression. The right panel shows how the compression affects the signal after it is normalized by its maximum amplitude.

Different amounts of glottal pulse spectral modification and signal compression was applied in building the voices according to the different SNR levels. Thus, three versions of the modified voices were created.

4. Results

The intelligibility of different systems were evaluated in the Hurricane Challenge listening test. In this paper, the results for the proposed systems (denoted as *GlottLombard*), natural utterances (denoted as *Plain*), and reference TTS system (denoted as *RefTTS*) are presented. Two noise types, both with three SNRs were used in the test as maskers: competing speech (CS) from a female talker at utterance-wise SNRs of -7, -14, and -21 dB, and speech-shaped noise (SSN) whose long-term average spectrum matched that of the CS, at SNRs of 1, -4, and -9 dB. The submitted sentences were normalized with respect to the

Table 2: Results of the subjective evaluation. Upper table shows the percent of correctly recognized words for each system in each condition. Data is represented as means and standard errors. Bottom table shows the intelligibility gain in decibels for the proposed system (GlottLombard) with respect to original speech (Plain) and the reference TTS system (RefTTS).

Masker	Competing speaker (CS)			Speech-shaped noise (SSN)		
	SNR = -7 dB	SNR = -14 dB	SNR = -21 dB	SNR = 1 dB	SNR = -4 dB	SNR = -9 dB
GlottLombard	59.4 $\pm$ 2.2 %	35.9 $\pm$ 1.5 %	17.1 $\pm$ 2.3 %	67.3 $\pm$ 2.0 %	52.0 $\pm$ 1.3 %	23.8 $\pm$ 2.2 %
Plain	85.1 $\pm$ 2.1 %	57.0 $\pm$ 2.4 %	24.8 $\pm$ 1.9 %	88.3 $\pm$ 2.1 %	63.0 $\pm$ 2.2 %	17.3 $\pm$ 2.1 %
RefTTS	59.7 $\pm$ 1.6 %	31.3 $\pm$ 1.9 %	11.7 $\pm$ 1.3 %	63.7 $\pm$ 1.9 %	32.8 $\pm$ 1.8 %	6.8 $\pm$ 1.2 %
Gain w.r.t. Plain	-7.39 dB	-4.67 dB	-2.54 dB	-3.83 dB	-1.33 dB	1.18 dB
Gain w.r.t. RefTTS	-0.06 dB	1.14 dB	2.42 dB	0.47 dB	2.34 dB	4.30 dB

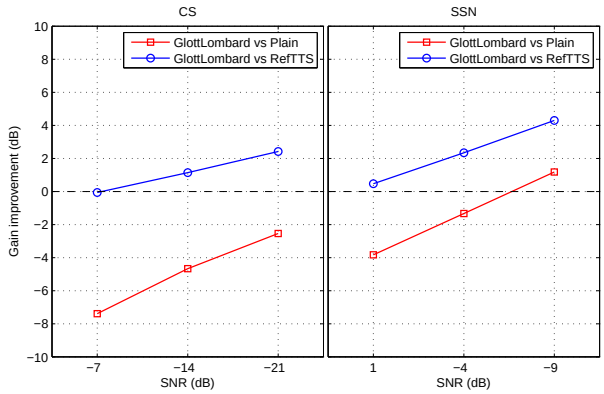

Figure 5: *Improvement in gain (dB) for the proposed system (GlottLombard) in comparison to original speech (Plain) and reference TTS system (RefTTS) in competing speech (CS) and speech-shaped noise (SSN) with three different SNRs.*

root mean square (RMS) energy and mixed with the maskers according to the SNRs. In the test, the listeners heard the sentence and masker through headphones only once, after which they typed what they heard. A total of 175 young adult listeners who passed an audiological screening and had no reported hearing disorders participated in the test.

The results of the intelligibility evaluation are shown in Table 2 and illustrated in Figure 4. The proposed method is significantly less intelligible than original speech, which is a rather expected result as synthetic voices tend to be less intelligible compared to natural speech [23]. However, the intelligibility difference between original speech and the proposed TTS voice becomes smaller as SNR is decreased, and in SSN condition with the SNR of -9 dB, the proposed system is more intelligible than original speech. Compared to the reference TTS voice, the proposed TTS voice is more intelligible in all conditions except for the high-SNR conditions of both noise types, in which cases the standard errors of the intelligibility ratings overlap each other.

The intelligibility gain in decibels for the proposed voice in comparison to original speech and the reference TTS voice is shown in Table 2 (bottom) and illustrated in Figure 5. In comparison to original speech, positive gain is achieved only with SSN masker with SNR of -9 dB, but compared to the reference TTS system, positive gains as high as 2–4 dB are achieved with low-SNR conditions.

5. Discussion

Although our system performed significantly better than the baseline TTS system, the results in this challenge were slightly inferior to our previous experiences on Lombard speech synthesis, where we have been able to achieve superior intelligibility compared to natural modal voice [5, 6]. Having considerable differences between setups, we may only speculate the reasons of the degraded performance. First, the simple voice conversion method has an obvious weakness of performing transformation per-utterance basis, causing distortion of formants if the distribution of phonemes in an utterance is atypical. This could be easily remedied by pre-calculating the statistics of the base voice from larger number of synthesized utterances. Second, noting the lower performance on competing female speaker condition, the raising of F0 may have had detrimental effect on the intelligibility because of increased overlap in the F0 range and thus added difficulty in following the flow of the target speech.

Finally, the uniform increase of duration by 10% had the effect of increasing the total RMS energy of each utterance. Thus, as the speech files were mixed with noise with respect to the total RMS energy in order to achieve specific SNRs for the test, the average level of the lengthened utterances were lower compared to speech without duration increase. This may have had a dramatic effect on the intelligibility of the proposed system, although this type of level normalization does not correspond to a real situation, where only the peak amplitude and maximum average frame-wise RMS energy is of concern.

6. Conclusions

This paper described modification of a TTS system for improving the intelligibility of speech in various noise conditions. The GlottHMM vocoder was used for training a voice with modal speech data, after which the voice parameters were modified in an unsupervised manner to mimic the properties of Lombard effect based on a small amount of Lombard speech from the same speaker. The evaluation results of the Hurricane Challenge 2013 show that the modified voice is more intelligible than unmodified natural speech only if masked by low-SNR speech-shaped noise, but equal or more intelligible compared to the reference TTS systems in all conditions.

7. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement n° 287678 and from the Academy of Finland (256961, 135003).

8. References

[1] Zen, H., Tokuda, K. and Black, A.W., "Statistical parametric speech synthesis", Speech Commun., 51(11):1039–1064, 2009.

[2] ITU, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs" International Telecommunication Union, Recommendation P.862, 2001.

[3] Cooke, M., "A glimpsing model of speech perception in noise", J. Acoust. Soc. Am., 119(3):1562-1573, 2006.

[4] Villegas, J. and Cooke, M., "Maximising objective speech intelligibility by local f0 modulation", Proc. Interspeech, 2012.

[5] Suni, A., Raitio, T., Vainio, M. and Alku, P., "The GlottHMM speech synthesis entry for Blizzard Challenge 2010", The Blizzard Challenge 2010 workshop, 2010. Online: http://festvox.org/blizzard

[6] Raitio, T., Suni, A., Vainio, M. and Alku, P., "Analysis of HMM-based Lombard speech synthesis", Proc. Interspeech, 2011, pp. 2781–2784.

[7] Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M. and Alku, P., "HMM-based speech synthesis utilizing glottal inverse filtering", IEEE Trans. on Audio, Speech, and Lang. Proc., 19(1):153–165, 2011.

[8] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.W. and Tokuda, K., "The HMM-based speech synthesis system (HTS) version 2.0", Sixth ISCA Workshop on Speech Synthesis, 2007, pp. 294–299.

[9] [Online] HMM-based speech synthesis system (HTS), *http://hts.sp.nitech.ac.jp*

[10] Suni, A., Raitio, T., Vainio, M. and Alku, P., "The GlottHMM entry for Blizzard Challenge 2011: Utilizing source unit selection in HMM-based speech synthesis for improved excitation generation", The Blizzard Challenge 2011 workshop, 2011. Online: http://festvox.org/blizzard

[11] Suni, A., Raitio, T., Vainio, M. and Alku, P., "The GlottHMM entry for Blizzard Challenge 2012 – Hybrid approach", The Blizzard Challenge 2012 workshop, 2012. Online: http://festvox.org/blizzard

[12] Raitio, T., Suni, A., Pulakka, H., Vainio, M. and Alku, P., "HMM-based Finnish text-to-speech system utilizing glottal inverse filtering", Proc. Interspeech, 2008, pp. 1881–1884.

[13] Raitio, T., Suni, A., Pulakka, H., Vainio, M. and Alku, P., "Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis" Proc. ICASSP, 2011, pp. 4564–4567.

[14] Alku, P., "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering", Speech Commun., 11(2–3):109–118, 1992.

[15] Alku, P., Tiitinen, H. and Näätänen, R., "A method for generating natural-sounding speech stimuli for cognitive brain research", Clinical Neurophysiology, 110:1329–1333, 1999.

[16] Soong, F.K. and Juang, B.-H., "Line spectrum pair (LSP) and speech data compression", Proc. ICASSP, vol. 9, 1984, pp. 37–40.

[17] Marume, M., Zen, H., Nankaku, Y., Tokuda, K. and Kitamura, T., "An investigation of spectral parameters for HMM-based speech synthesis", Proc. Autumn Meeting of Acoust. Soc. of Japan, 2006 (In Japanese).

[18] Raitio, T., Suni, A., Vainio, M. and Alku, P., "Comparing glottal-flow-excited statistical parametric speech synthesis methods", accepted for publication in Proc. ICASSP, Vancouver, Canada, May 26–31, 2013.

[19] Qian, Y., Soong, F.K., Chen, Y., Chu, M., "An HMM-based Mandarin Chinese text-to-speech system", ISCSLP, 2006, pp. 223–232.

[20] Yamagishi, J. and Watts, O., "The CSTR/EMIME HTS System for Blizzard Challenge 2010", The Blizzard Challenge 2010 workshop, 2010. Online: http://festvox.org/blizzard

[21] [Online] Festival Speech Synthesis System, *http://festvox.org/festival/*

[22] Raitio, T., Suni, A., Pulakka, H., Vainio, M. and Alku, P., "Comparison of formant enhancement methods for HMM-based speech synthesis", Seventh ISCA Workshop on Speech Synthesis, 2010, pp. 334–339.

[23] King, S. and Karaiskos, V., "The Blizzard Challenge 2010", The Blizzard Challenge 2010 workshop, 2010. Online: http://festvox.org/blizzard