INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Using text and acoustic features in predicting glottal excitation waveforms for parametric speech synthesis with recurrent neural networks

Lauri Juvela[1,2], *Xin Wang*[2,3], *Shinji Takaki*[2], *Manu Airaksinen*[1], *Junichi Yamagishi*[2,3,4], *Paavo Alku*[1]

[1]Aalto University, Department of Signal Processing and Acoustics, Finland
[2]National Institute of Informatics, Japan
[3]Sokendai University, Japan
[4]University of Edinburgh, The Centre for Speech Technology Research, United Kingdom
{lauri.juvela,manu.airaksinen,paavo.alku}@aalto.fi, {wangxin,takaki,jyamagis}@nii.ac.jp

Abstract

This work studies the use of deep learning methods to directly model glottal excitation waveforms from context dependent text features in a text-to-speech synthesis system. Glottal vocoding is integrated into a deep neural network-based text-to-speech framework where text and acoustic features can be flexibly used as both network inputs or outputs. Long short-term memory recurrent neural networks are utilised in two stages: first, in mapping text features to acoustic features and second, in predicting glottal waveforms from the text and/or acoustic features. Results show that using the text features directly yields similar quality to the prediction of the excitation from acoustic features, both outperforming a baseline system based on using a fixed glottal pulse for excitation generation.

Index Terms: parametric speech synthesis, glottal vocoding, excitation modelling, LSTM

1. Introduction

Statistical parametric speech synthesis (SPSS) [1, 2] has become a widely used speech synthesis technique in recent years. The statistical approach has several attractive properties, including good generalisation on unseen text inputs, flexible speaker adaptation and a small memory footprint compared to unit selection synthesis [3]. However, the overall quality of parametric synthesis has not yet reached that of the best unit selection techniques. Recently, deep learning techniques have been successfully implemented in the acoustic modelling for SPSS [4, 5]. Further improvement has been attained from the use of recurrent neural networks (RNNs), taking advantage of the sequential nature of the parametric speech representation, specifically by using long short-term memory (LSTM) networks [6] and bidirectional LSTM [7]. Despite these advances in the acoustic modelling, the resulting synthetic speech quality still depends on the underlying parameterisation and reconstruction of the speech signal, a process known as vocoding.

Vocoders are typically based on the source-filter model of speech, where a filter conveying speech spectral information is excited with a source signal. The most prevalent vocoder in SPSS uses a STRAIGHT [8, 9]-based mel-generalised cepstrum (MGC) representation for the filter part together with a mixed excitation signal created by modifying an impulse train to satisfy a specific band-aperiodicity measure. However, using an impulse train excitation results in a perceptual degradation described as "buzziness" in the reconstructed speech, due to too much high frequency energy and the zero-phase characteristic of the impulse. How to best generate a natural excitation signal and phase for speech synthesised from a parametric representation remains an open research question.

Several approaches have been proposed to create more natural vocoded speech. Vocoders relying on the source-filter model focus on improving the excitation model: proposed techniques include the deterministic plus stochastic model (DSM) [10], which uses principal component analysis on the filter residuals to create eigen-representations of residual waveforms, an MGC-based vocoder replacing the impulse train with the Liljencrants-Fant parametric model for glottal excitation [11], and the GlottHMM vocoder [12], which uses glottal inverse filtering (GIF) for vocal tract filter estimation and a natural glottal flow pulse for creating the excitation. There are also some hybrid approaches that use statistical modelling for the acoustic parameters and unit-selection for excitation generation with a residual codebook [13] or a glottal pulse library [14]. Similarly to unit selection, the hybrid approach faces the difficulty of selecting the best unit in terms of acoustic concatenation criteria. In addition to the source-filter-based vocoders, another approach is the use of sinusoidal vocoders [15, 16] that create harmonic sinusoidal components based on the spectral envelope and fundamental frequency (f_0) information. These methods encounter the problem of "inventing the phase" as well, and typically use a minimum phase derived from magnitude spectrum, which is not entirely justified from the voice production perspective. An experimental comparison of different vocoder types found that the sinusoidal vocoders suitable for SPSS have comparable quality to the source-filter vocoders in an analysis-synthesis setup [17].

Recently, deep neural networks (DNN) have been applied in modelling the glottal pulse waveforms, increasing the overall quality and flexibility for varying vocal effort [18, 19]. The glottal flow derivative waveforms are first estimated by the iterative adaptive inverse filtering (IAIF) [20] technique and then a neural network is trained to predict these waveforms from the other acoustic features. However, this approach is somewhat sensitive to the accuracy of GIF and glottal closure instant (GCI) detection. More recently, improved synthesis quality was reported in [21] for a high-pitched voice by using a more advanced GIF method, the quasi-closed phase (QCP) [22] inverse filtering. Overall, taking this kind of modelling approach provides increased dynamics for the voice source model in a data driven manner while overcoming the problems with pulse selection in the hybrid approach.

Previous TTS systems utilising glottal vocoding have used

http://dx.doi.org/10.21437/Interspeech.2016-712

HMM-based acoustic models mixed together with a DNN-based excitation model [18, 19, 21]. Compared to these earlier glottal synthesis systems, the current study has three novelties: 1) the HMM-based acoustic models are replaced by deep bidirectional LSTM. 2) the architecture for the DNN excitation model is also changed from FF to RNN, and 3) we further investigate various inputs for the LSTM-based excitation modelling, including acoustic and text features.

The paper is structured as follows: in section 2 we overview the synthesis system structure with the various options for excitation modelling, with subsections covering the acoustic features, text features, and glottal waveform formatting for DNN. Section 3 details the experiments on training the synthesis systems and subjective evaluation of the investigated excitation methods.

2. Speech synthesis system

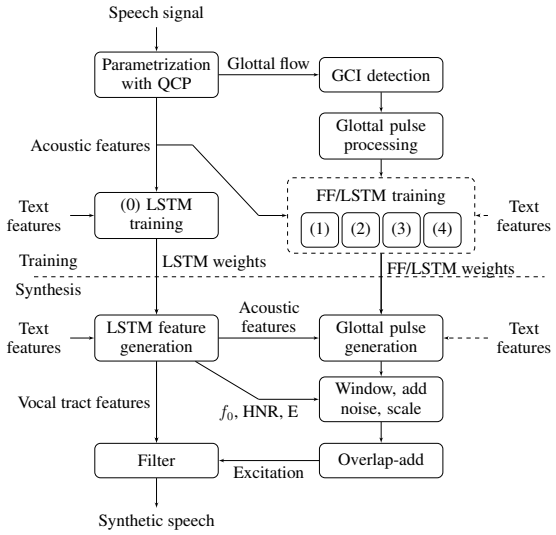

Figure 1: *Overview of the speech synthesis system. Four different networks with feedforward or LSTM structure, and text and/or acoustic feature inputs are used for modelling the glottal excitation waveforms.*

The TTS system examined here is based on our recent HMM-based platform that utilises glottal vocoding [21]. While the acoustic parameterisation and the processing of glottal waveforms remains unchanged, our new contribution in this work is the investigation of different modelling techniques. First, the acoustic model of the TTS system now utilises deep bidirectional LSTM instead of the previous HMM-based approach. Second, we investigate the prediction of the glottal excitation waveforms using various configurations based on deep learning. Figure 1 shows the general structure of the proposed synthesis system. At the training stage, the acoustic parameters, including vocal tract and glottal source features, are first estimated from the speech signal frame-wise using QCP glottal inverse filtering, as described in section 2.1. The acoustic features are aligned with the text features (see section 2.2) to train the base synthesis network mapping the text to acoustic features. For the excitation modelling, glottal pulses are extracted from the inverse filtering result and processed, as described in section 2.3. At the synthesis stage, acoustic parameters are generated from given text using the LSTM network, from where they are fed into the excitation modelling networks. Alternatively, text features are used directly to generate the excitation pulses. After generation, the pulses are truncated and windowed in accordance with f_0, modified for aperiodicity in accordance with the harmonic-to-noise ratio (HNR), and scaled to the desired energy level. Finally the excitation signal is created with overlap-add and filtered in accordance with the generated vocal tract filter.

The different deep learning-based systems used in this work are summarised in Table 1. To focus the comparison on the excitation models, a base synthesis system mapping text features to acoustic features (TXT-AC) is shared among the systems. Four different systems for generating the glottal excitation are trained. (1) A system using a feedforward (FF) network to map acoustic features to glottal pulses (AC-GL-FF) is conceptually equal to that in [21], and can be considered the baseline deep learning excitation model. (2) By replacing the feedforward network in AC-FF-GL with LSTM, we obtain a new system called AC-LSTM-GL. Using a recurrent network can be hypothesised to add context awareness to the network, potentially improving performance at phoneme boundaries or phonation onsets. Another novel concept in this work is introducing the text features into excitation modelling. To test whether it is possible to predict glottal waveforms using only the text information, we build a new network (3) to map the text features directly to the glottal pulses (TXT-LSTM-GL). Furthermore, the previous approach of predicting from acoustic features could benefit from additional context-dependent linguistic information. To test this, we create the final network (4) by concatenating the acoustic and text features to the network's input and train the network (TXT+AC-LSTM-GL) to map this information to the glottal pulses. Both of the methods utilising text features use an LSTM network with the same internal topology.

Table 1: *In total, five systems were trained for the experiments: The base synthesis system (0) generating acoustic features (AC) from text features (TXT) is shared among compared systems. The compared systems (1–4) generate glottal excitation waveforms (GL) from text and/or acoustic features as follows:*

ID	System	Input	Output	Network
(0)	TXT-LSTM-AC	TXT	AC	LSTM
(1)	AC-FF-GL	AC	GL	FF
(2)	AC-LSTM-GL	AC	GL	LSTM
(3)	TXT-LSTM-GL	TXT	GL	LSTM
(4)	TXT+AC-LSTM-GL	TXT + AC	GL	LSTM

2.1. Acoustic features

The extraction of acoustic features is performed similarly to [21]. First, the speech signal is analysed with the QCP inverse filtering, giving estimates for the vocal tract filter and the glottal source. The vocal tract is represented by an all-pole filter whose LSF coefficients (LSF VT) are used as the parameterisation. Additional parameters are estimated from the glottal source, i.e., the source spectral envelope is represented by all-pole filter LSF parameters (LSF SRC). The fundamental frequency f_0 and the voiced-unvoiced decision (VUV) are further estimated from the glottal source. Finally, the harmonic-to-noise ratio (HNR) of the glottal source is estimated to measure aperiodicity in the excitation. The signal frame energy is included as well for scaling in the synthesis stage. For neural network modelling, the

acoustic features and their dynamics (Δ and $\Delta\Delta$) are used. The f_0 is modelled continuously and a separate binary VUV feature is added. Table 2 lists the acoustic parameters and their dimensions as used in the DNNs.

Table 2: Acoustic features and their dimensions (including their Δ and $\Delta\Delta$ values) used in the various DNN models. The f_0 and VUV are modelled separately in the DNN while the f_0 vector otherwise contains the voicing information.

Feature	dim.	Δ dim.
f_0	1	3
VUV	–	1
Energy	1	3
LSF VT	30	90
LSF SRC	10	30
HNR	5	15
total	47	142

2.2. Text features

The same set of context-dependent linguistic features, called the text features for short, is used for predicting both the acoustic features and the glottal waveforms. These text features, which include phoneme, syllable, word, phrase, and sentence level information, are generated from the text with the Flite [23] speech synthesis front-end using the Combilex [24] US English lexicon, resulting in a total dimensionality of 396. An alignment between the phoneme-rate text features and the frame-rate signal features is found with the HMM-based speech synthesis system (HTS) [25], and at the synthesis stage, the HTS duration models are used to create the input text features to the DNNs.

2.3. Glottal waveforms

The glottal waveforms estimated by GIF are processed for the deep network training similarly to [21], as shown in Fig. 2: take a two pitch-period segment from the estimated glottal volume velocity derivative waveform, having glottal closure instants (GCI) at the middle and at both ends, apply cosine windowing, and zero-pad the pulse symmetrically to match the fixed network output dimension of 400 samples. At the synthesis stage, the generated pulses are truncated to match the desired f_0, windowed again to compete the squared cosine windowing, and overlap-added to create the excitation. Additionally, for modelling with RNN the glottal waveforms must be sequential. In this work, we associate one glottal pulse with each voiced frame by taking the nearest pulse. A zero-vector is associated with unvoiced frames.

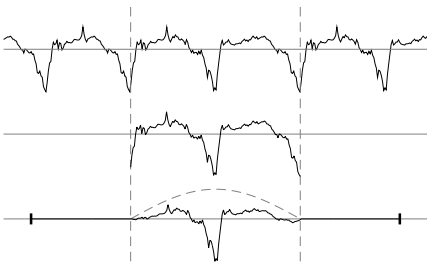

Figure 2: *Processing of the glottal flow derivative waveforms for training the networks: a two pitch-period segment delimited by GCI is cosine windowed and zero-padded to desired network output length.*

3. Experiments

3.1. Speech material

The speech material used for training the synthesiser was produced by a female US English speaking professional voice talent [26]. The dataset comprises approximately 12,000 utterances totalling 14 hours, of which 500 utterances were used as a validation set in training and 200 were kept as an unseen test set for generation, while using the rest for training. The speech was downsampled to 16 kHz sample rate from the original 48 kHz.

3.2. Training the synthesis systems

All the networks had four hidden layers: for the feedforward network, the hidden layers were of size 512 with logistic activation functions, while the LSTM networks consisted of two feedforward logistic hidden layers of size 512 with two bidirectional LSTM layers of size 256 stacked on top of them. The CURRENNT toolkit [27] was employed in training the networks. For all networks, the training was stopped after 5 epochs of no improvement on the validation set. The sum of squared error (SSE) for the training and validation sets is presented for each method in Fig. 3 as a function of training epochs, where the solid lines correspond to training error and dashed lines to validation error. The error measures show that the text-only network has higher error level and starts overfitting early compared to the networks including acoustic inputs. Fig. 4 shows an example of generated excitation waveforms from the various systems without the added voiced excitation noise component.

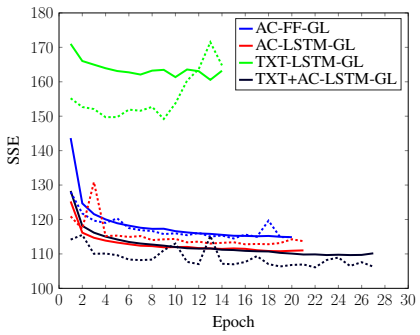

Figure 3: *The training errors (solid line) and validation errors (dashed line) for the DNN excitation systems.*

3.3. Subjective listening test

The test stimuli were created by first using the TTS front-end with HTS-based duration models to create the neural network text features and then inputting these to the TXT-LSTM-AC network to generate acoustic features before finally using the various excitation model networks to generate glottal pulses from the text and generated acoustic features.

A subjective listening test similar to the multiple stimulus test with hidden reference and anchor (MUSHRA) [28, 29] was performed, using a real speech sample from the target speaker with the same linguistic content as the reference. No low-quality anchor was included in the form of a degraded reference sample, since the degraded anchor would still have perfect timing and prosody, and comparing this with TTS samples is problematic. Instead, a single pulse excitation method, as used in the original GlottHMM [12], was included in the test to serve as a non-DNN baseline for the various DNN based excitation

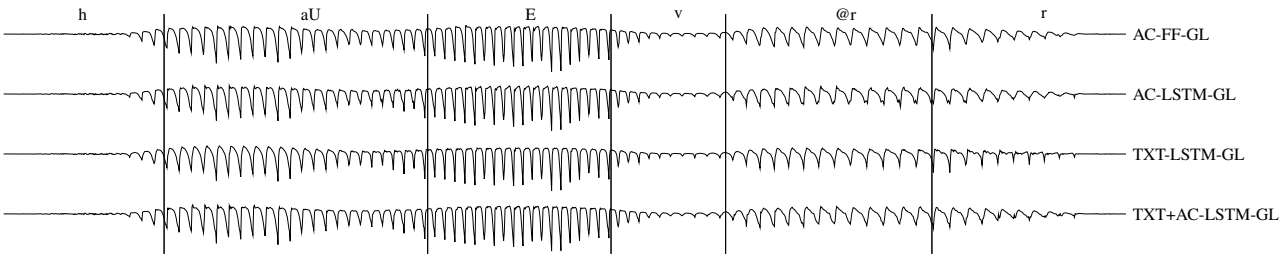

Figure 4: *Generated glottal derivative waveforms prior to adding the voiced excitation noise component. The phoneme boundaries are included to show how the waveforms change their shape along with the linguistic context. In this example, the target word is "however".*

methods. The evaluation was conducted in a listening booth environment using Beyerdynamic DT 990 headphones. Thirty native English speakers with no reported hearing disorders participated in the listening test, four of whom were excluded in post-screening due to inconsistency in finding the hidden reference or insufficient variance in their answers. The results were analysed with a repeated measures ANOVA [30] using Greenhouse–Geisser correction. Analysis shows that the overall test main effects of method $[F(2.94, 73.4) = 29.192, p < .001]$, and sample $[F(9.89, 247.2) = 10.118, p < .001]$, as well as the interaction method$\times$sample $[F(9.99, 249.9) = 2.134, p = .023]$ are statistically significant. Fig. 5 shows the estimated marginal means and 95% confidence intervals with Bonferroni correction. Post-hoc tests showed that the DNN-based excitation methods do not differ significantly, regardless of TXT-LSTM-GL having a slightly lower mean score. However, the lower rating of the single pulse method is statistically significant.

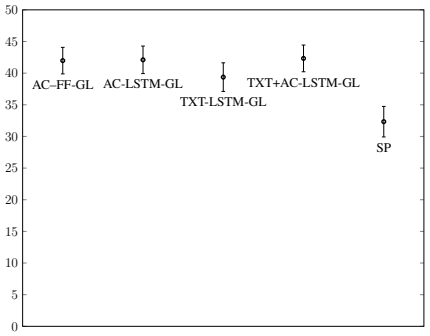

Figure 5: *Marginal means and 95% confidence intervals for the methods in the MUSHRA testing. The differences between DNN-based methods are not statistically significant, regardless of TXT-LSTM-GL scoring slightly lower. However, the score for the single pulse excitation (SP) is significantly lower.*

Since the MUSHRA test did not provide the resolution to differentiate between the top three methods, an additional preference test with forced choice was conducted for AC-FF-GL, AC-LSTM-GL and TXT+AC-LSTM-GL . The preference scores are presented in Fig. 6 with 95% confidence intervals estimated by normal approximation. Binomial tests indicate that the LSTM-based AC-LSTM-GL and TXT-AC-LSTM-GL were preferred over the feedforward AC-FF-GL with $p = .002$ and $p = .036$, respectively. Between the LSTM-based methods, AC-LSTM-GL was preferred with $p = .007$.

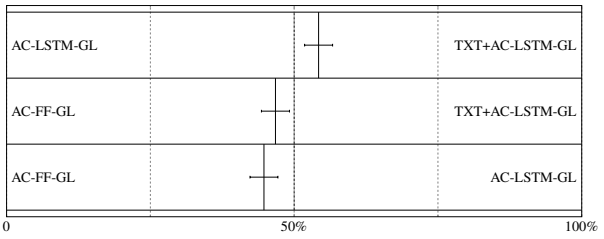

Figure 6: *The preference test shows that the LSTM-based methods are preferred over the feedforward method, while the LSTM network using only acoustic features outperforms the network using the concatenated text and acoustic input.*

4. Discussion and Conclusion

Our parametric speech synthesis experiments show that the glottal vocoder excitations can be predicted relatively well using text features, which implies that the linguistic context carries meaningful information about the voice source. While the text-to-glottal (TXT-LSTM-GL) system was rated slightly lower than the other deep learning based excitation systems, the system was still rated higher than the single pulse baseline. Moreover, the preference test indicated the LSTM-based excitation models outperforming the feedforward one. Replicating the experiments with a male voice could yield larger perceptual differences since the excitation phase captured in the waveform becomes more relevant with low-pitched voices.

The straightforward approach of using all available context information is likely not optimal, as all of the full context might not be relevant to the excitation, while the increased dimensionality makes the modelling problem more challenging. This is reflected in the preference test, as using only the acoustic features was preferred to the concatenated acoustic and text features. Selecting the most useful text features remains a task for future research. Other future work includes joint deep learning based modelling of the acoustic features and glottal waveforms aiming to better capture the interactions taking place, and attention modelling to disregard unvoiced regions in the voiced excitation model.

5. Acknowledgements

This work was supported by the Academy of Finland (proj. no. 256961 and 284671), the European Union TEAM-MUNDUS scholarship (TEAM1400081), and the EPSRC through Programme Grant EP/I031022/1 (NST) and EP/J002526/1 (CAF).

6. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. of Interspeech*, 1999, pp. 2347–2350.

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

[3] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

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

[5] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *Signal Processing Magazine, IEEE*, vol. 32, no. 3, pp. 35–52, May 2015.

[6] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4470–4474.

[7] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks." in *Interspeech*, 2014, pp. 1964–1968.

[8] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[9] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," in *MAVEBA*, 2001.

[10] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 968–981, March 2012.

[11] J. P. Cabral, S. Renals, J. Yamagishi, and K. Richmond, "HMM-based speech synthesiser using the LF-model of the glottal source," in *Proc. of ICASSP*. IEEE, 2011, pp. 4704–4707.

[12] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 153–165, January 2011.

[13] T. Drugman, A. Moinet, T. Dutoit, and G. Wilfart, "Using a pitch-synchronous residual codebook for hybrid HMM/frame selection speech synthesis," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, April 2009, pp. 3793–3796.

[14] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 4564–4567.

[15] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 1, pp. 21–29, 2001.

[16] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 184–194, April 2014.

[17] Q. Hu, K. Richmond, J. Yamagishi, and J. Latorre, "An experimental comparison of multiple vocoder types," in *8th ISCA Workshop on Speech Synthesis, Barcelona, Spain*, 2013, pp. 155–160.

[18] T. Raitio, H. Lu, J. Kane, A. Suni, M. Vainio, S. King, and P. Alku, "Voice source modelling using deep neural networks for statistical parametric speech synthesis," in *22nd European Signal Processing Conference (EUSIPCO)*, Lisbon, Portugal, September 2014.

[19] T. Raitio, A. Suni, L. Juvela, M. Vainio, and P. Alku, "Deep neural network based trainable voice source model for synthesis of speech with varying vocal effort," in *Proc. of Interspeech*, Singapore, September 2014, pp. 1969–1973.

[20] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2–3, pp. 109–118, 1992, Eurospeech '91.

[21] L. Juvela, B. Bollepalli, M. Airaksinen, and P. Alku, "High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network," in *"Proc. of ICASSP"*, Mar. 2016, pp. 5120–5124.

[22] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 3, pp. 596–607, March 2014.

[23] A. W. Black and K. A. Lenzo, "Flite: a small fast run-time synthesis engine," in *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, 2001.

[24] K. Richmond, R. A. Clark, and S. Fitt, "Robust LTS rules with the Combilex speech technology lexicon," in *Proc. of Interspeech*, Brighton, September 2009, pp. 1295–1298.

[25] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system version 2.0," in *Proc. of ISCA SSW6*, Bonn, Germany, August 2007, pp. 294–299.

[26] S. King and V. Karaiskos, "The Blizzard Challenge 2011," in *Blizzard Challenge 2011 Workshop*, Turin, Italy, September 2011.

[27] F. Weninger, "Introducing CURRENNT: The Munich open-source CUDA recurrent neural network toolkit," *Journal of Machine Learning Research*, vol. 16, pp. 547–551, 2015. [Online]. Available: http://jmlr.org/papers/v16/weninger15a.html

[28] "ITU-R recommendation BS.1534 : Method for the subjective assessment of intermediate quality levels of coding systems," International Telecommunication Union, Tech. Rep., 2015.

[29] E. Vincent, "MUSHRAM: A MATLAB interface for MUSHRA listening tests," http://c4dm.eecs.qmul.ac.uk/downloads/#mushram, 2005.

[30] H. J. Keselman, J. Algina, and R. K. Kowalchuk, "The analysis of repeated measures designs: a review," *British Journal of Mathematical and Statistical Psychology*, vol. 54, no. 1, pp. 1–20, 2001.