HIGH-PITCHED EXCITATION GENERATION FOR GLOTTAL VOCODING IN STATISTICAL PARAMETRIC SPEECH SYNTHESIS USING A DEEP NEURAL NETWORK

Lauri Juvela, Bajibabu Bollepalli, Manu Airaksinen, Paavo Alku

Aalto University, Department of Signal Processing and Acoustics, Finland

ABSTRACT

Achieving high quality and naturalness in statistical parametric synthesis of female voices remains to be difficult despite recent advances in the study area. Vocoding is one such key element in all statistical speech synthesizers that is known to affect the synthesis quality and naturalness. The present study focuses on a special type of vocoding, glottal vocoders, which aim to parameterize speech based on modelling the real excitation of (voiced) speech, the glottal flow. More specifically, we compare three different glottal vocoders by aiming at improved synthesis naturalness of female voices. Two of the vocoders are previously known, both utilizing an old glottal inverse filtering (GIF) method in estimating the glottal flow. The third on, denoted as Quasi Closed Phase – Deep Neural Net (QCP-DNN), takes advantage of a recently proposed new GIF method that shows improved accuracy in estimating the glottal flow from high-pitched speech. Subjective listening tests conducted on an US English female voice show that the proposed QCP-DNN method gives significant improvement in synthetic naturalness compared to the two previously developed glottal vocoders.

Index Terms— Statistical parametric speech synthesis, Glottal vocoder, Deep neural network, Glottal inverse filtering, QCP

1. INTRODUCTION

Statistical parametric speech synthesis, or HMM-based synthesis [1, 2], has become a popular speech synthesis technique in recent years. The benefits of the framework include flexible voice adaptation, robustness and small memory footprint. In general, however, statistical speech synthesis methods are not capable of yielding as good speech quality as the best unit selection techniques. This stems mainly from three causes [2, 3]: First, the parametric representation of speech, the process called vocoding, is unable to represent the speech waveform adequately hence resulting in robotic quality and buzziness. Second, HMMs generate over-smoothed parameters due to statistical averaging which results in a muffled voice character. Finally, there are inaccuracies in the statistical acoustic modelling, where the dynamic model produces smooth parameter trajectories, causing additional muffling, particularly at phone transitions. Despite recent advances in the acoustic modelling with deep neural networks (DNNs) [4, 5], the statistical speech synthesis paradigm still relies on the underlying speech parametrization. Therefore, improved speech parametrization by using more advanced vocoding techniques constitutes a justified topic when aiming at better quality and naturalness of synthetic speech.

The source-filter model is a widely used parametric representation of speech. In traditional source-tract models, the spectral envelope of speech is captured by a linear prediction (LP) synthesis filter and the signal is synthesized using a spectrally flat excitation (impulse train or noise). Using this kind of overly simplified excitation waveforms in vocoding, however, is likely the cause of the distinctive buzziness in statistical parametric speech synthesis. The most widely used vocoder, STRAIGHT [6, 7], attempts to tackle this problem by adding noise-like aperiodicity into the impulse train excitation hence breaking its zero-phase characteristic. The excitation phase information has been shown to affect the synthetic speech quality [8, 9] and therefore further attention should be directed to the excitation signal at waveform level.

As an alternative to overly simplified excitation waveforms, a vocoding approach based on modelling the real excitation of human speech production, the glottal flow, was introduced in [10]. This vocoder, named GlottHMM, takes advantage of glottal inverse filtering (GIF) in order separate the speech signal into a glottal flow and vocal tract in the training phase of the statistical synthesis. In the synthesis part, the vocoder reconstructs the speech waveform by using a glottal flow pulse, called the library pulse, that has been estimated in advance from natural speech, and a set of acoustical parameters obtained from HMMs. Subjective listening tests on a male Finnish voice in [11] indicated that the speech quality obtained with GlottHMM was superior to that produced by STRAIGHT. In addition, the glottal based vocoding approach was shown to be the most successful technique in Blizzard Challenge 2010 [12] in experiments where intelligibility of synthetic speech was assessed in noisy conditions: the GlottHMM enabled adapting the speaking style according to the natural Lombard effect hence achieving the best score in intelligibility tests. Recently, a new version of GlottHMM was proposed based on combining a HMM-based synthesis system with a glottal vocoder which uses DNNs instead of pre-computed library pulses in generation of the excitation waveform [13]. Subjective listening experiments reported in [13, 14] indicate that the DNN-based generation of the vocoder excitation resulted in a small, yet significant quality improvement.

Despite recent advances both in statistical mapping (i.e. replacing HMM-based platforms with DNN-based ones) and in vocoding, naturalness of statistical synthesis still lags behind that of real speech. In particular, several studies (e.g. [15, 16]) have reported lower evaluation scores for synthetic female voices than for male voices. Therefore, there is a great need for better synthesis techniques capable of improving naturalness of high-pitched female voices. Vocoding, either as a part of a HMM-based or DNN-based synthesis platform, is undoubtedly one such key component that calls for new research when aiming at high quality synthesis of female speech. Given the fact that the glottal vocoding approach has succeeded in improving synthesis quality of male speech in a few recent years, as reported above, the present study was launched to examine whether this improvement can be achieved also for female voices. The study is motivated not only by the general need for better statistical synthesis techniques capable of generating high

This research was supported by the Academy of Finland, project nos. 256961 and 284671

quality female voices, but also by our recent advances in GIF techniques that show improved estimation accuracy in the computation of glottal flow from high-pitched speech [17]. The study compares three glottal based vocoders: the baseline GlottHMM introduced in [10], the DNN-based estimation of the excitation developed in [13] and the new one proposed in this study. The evaluation shows that the proposed method gives significant quality improvement for the synthetic speech of the tested female voice.

2. COMPUTATION OF THE VOCODER EXCITATION

The three vocoders to be evaluated are all based on the utilization of GIF in speech parametrization. The vocoders are different particularly with respect to how the excitation waveform in the synthesis stage is formed. In the following two sub-sections, the excitation modelling in these three vocoders is discussed by first shortly describing in section 2.1 the baseline and the current DNN-based technique, after which the proposed new DNN-based excitation modelling approach is described in detail in section 2.2.

2.1. Reference methods

The baseline of our comparison is the GlottHMM vocoder [11]. The method uses Iterative Adaptive Inverse Filtering (IAIF) [18] as the GIF method to separate the voice source and the vocal tract. Excitation waveform is computed in the synthesis stage from a single glottal flow library pulse that is estimated in advance from natural speech. The excitation pulse is modified to the desired pitch, source spectral tilt and harmonic-to-noise ratio (HNR), after which the concatenated excitation is filtered with the vocal tract filter to synthesize speech. In the rest of this paper, this method is referred to as **IAIF baseline**.

The current version of our statistical synthesizer uses a DNN-based voice source generation method introduced recently in [13]. The method is based on replacing the pre-computed library pulse used in IAIF baseline with a DNN based estimation of the excitation waveform. The DNN is trained to estimate the glottal flow computed by IAIF from a given acoustical feature vector. In the synthesis stage, the DNN generates a pitch and energy normalized excitation waveform for the vocoder. In the present study, this method is referred to as **IAIF-DNN**. The new method proposed in this study involves several computational blocks that were present already in [13]. These differences are clarified in describing the new method in section 2.2.

Our previous studies [13, 14] indicate that IAIF-DNN yields a small improvement in quality and naturalness compared to IAIF baseline. In addition, IAIF-DNN benefits from a more flexible control of the speaking style [14]. Quality improvements achieved with IAIF-DNN in [13, 14] were, however, smaller than expected. Most evident reasons why the use of DNNs in our previous experiments did not show a larger subjective quality enhancement are as follows: First, while a feature-to-waveform mapping by DNN succeeds in modelling the overall glottal flow waveform structure, it unfortunately also generates averaging which is manifested in the loss of finer high-frequency components of the vocoder excitation. Second, IAIF-DNN takes advantage of interpolation in normalizing the pitch of the glottal flow excitation. This type of pitch normalization causes additional high-frequency loss, as interpolators effectively act as low-pass filters [12]. This phenomenon is particularly detrimental for higher-pitched voices where the effect of pitch modification is stronger. Finally, and perhaps most importantly, the IAIF method, as many older GIF methods, is known to have poor accuracy in estimating glottal flows from voices of high pitch [19, 20]. For high-pitched

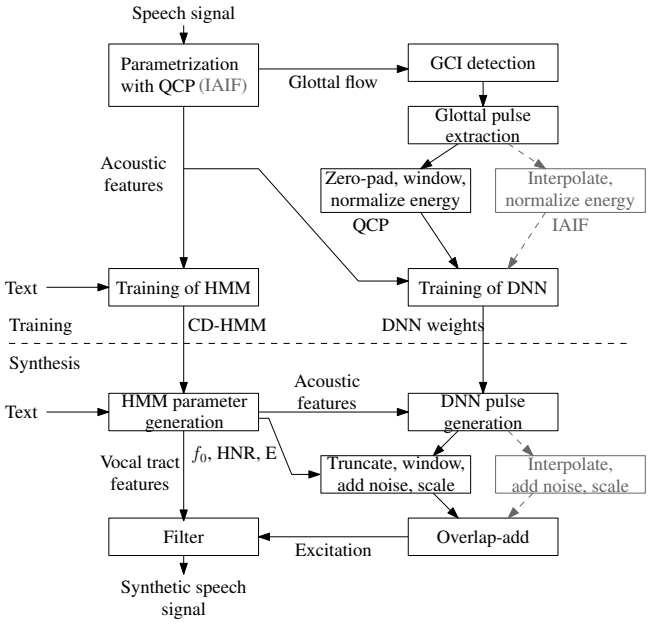


Fig. 1. Block diagram of the IAIF-DNN and QCP-DNN synthesis systems. Blocks corresponding to IAIF-DNN are drawn in grey.

speech, performance of the all-pole models used in older GIF methods deteriorates in estimating the vocal tract due to the contribution of sparse harmonics in the speech spectrum [21, 22]. Consequently, the estimated time-domain glottal excitation is degraded by incorrectly cancelled resonances of the tract. This poor separation of the speech waveform into the glottal flow and vocal tract in turn leads to degraded statistical modelling of the corresponding parameter distributions by HMMs which, finally, hinders achieving larger improvements in synthesis quality and naturalness.

2.2. Proposed method

The main modification in vocoding method proposed in the present study is that it uses a new GIF method, Quasi Closed Phase (QCP), which has been shown to perform well with high-pitched speech [17]. The block diagram of the proposed statistical synthesis system utilizing QCP is presented in Fig. 1. This new DNN-based method to compute the vocoder excitation is referred to as **QCP-DNN**.

In both IAIF-DNN and QCP-DNN, the DNN is trained with the GlottHMM feature vectors as the input and the vocoder excitation (i.e the time-domain glottal flow) as the output. Additionally, the output target waveforms in both IAIF-DNN and QCP-DNN consist of two consecutive glottal flow derivative pulses where glottal closure instants (GCIs) are located at both ends and in the middle of the two-cycle segment. There are three main differences between IAIF-DNN and QCP-DNN: first, as mentioned above, the former takes advantage of IAIF in the estimation of the glottal flow while the latter is based on QCP. Second, the target waveforms are treated differently: In IAIF-DNN, the output vectors are interpolated to cover a constant span of 400 samples regardless of the underlying fundamental frequency (f_0), and the energy is normalized. Hann windowing is used on the output waveforms to enable the use of overlap-add (OLA) for synthesizing the excitation from the generated pulses. In QCP-DNN, however, the interpolation is not used but the DNN is trained in such a way that it enables directly generating the excitation waveform of

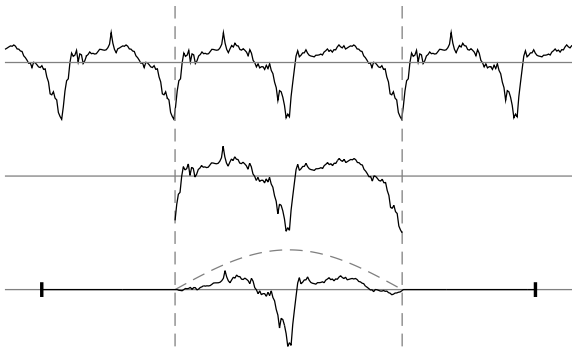


Fig. 2. To create a QCP-DNN output vector (bottom), a two-pitch-period segment (middle) is extracted from the glottal flow derivative waveform (top), cosine windowed and zero-padded to desired length. Respective zero-levels of the time domain waveforms are represented by horizontal lines.

a given pitch. This was achieved by changing the IAIF-DNN training so that the target waveforms are not interpolated, but are rather symmetrically zero padded to match the desired output length. The process is illustrated in Fig. 2. Moreover, the Hann, or squared cosine, windowing required for the OLA synthesis is broken into two cosine windowing parts: first before training and second time after generating the waveform from the DNN. This procedure eliminates any discontinuities caused by truncating the generated waveform to pitch period length. Finally, QCP-DNN uses the SEDREAMS GCI detection algorithm [23], which has been shown to perform well with speakers with various f_0 ranges [24], instead of the previously used IAIF residual based method. The need for accurate GCI detection is two-fold: the QCP inverse filtering algorithm requires reliable GCI estimates to achieve best results, and the GCIs are used in extracting the pulse waveforms for training.

3. TRAINING THE SYNTHESIS SYSTEMS

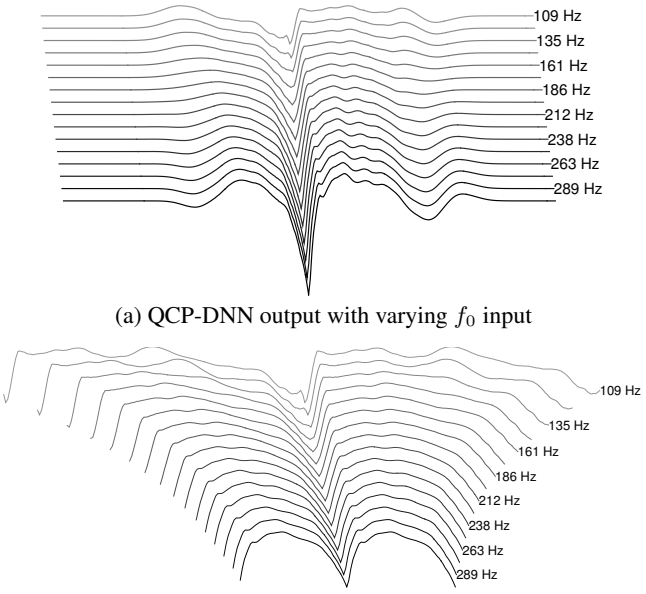
3.1. Speech material

In the experiment, we used the SLT-speaker from the CMU ARCTIC database [25] sampled at 16 kHz. The speaker is an U.S. English professional speaker commonly used in, for example, HTS speech synthesis demonstrations. The entire speech dataset consists of 1132 utterances, 60 of which were reserved for testing and the rest were used for training the speech synthesis system. The dataset is provided with context dependent phonetic labels with time alignment, which we used in training the HMM synthesis system.

3.2. Training of the DNNs

The DNN used in [13] was a standard feed-forward multilayer perceptron with sigmoid activation functions, random initialization and MSE-backpropagation training. In this study, we use the same network structure for both IAIF-DNN and QCP-DNN in order to focus on differences between the inverse filtering techniques. However, we modified the QCP-DNN error criterion to emphasize the main excitation peak of the glottal flow derivative waveform to better retain the high-frequency information carried by the peak.

In the experiments, two different DNN systems were trained: IAIF-DNN and QCP-DNN. Both systems are speaker dependent and the training data for the methods was derived from the same subset



(a) QCP-DNN output with varying f_0 input

(b) Overlap-added waveform

Fig. 3. QCP-DNN generated pulses with varying the f_0 at DNN input while keeping other parameters constant. The resulting overlap-added two-pitch-cycle waveform shows the effect more clearly.

of the SLT-speaker speech. An identical network topology was selected for both methods: A fully connected feed-forward multilayer perceptron with three hidden layers, sigmoid activation functions, and random initial weights drawn from the Gaussian distribution. The layer sizes were 47 for input, 100, 200, and 300 for the hidden layers, and the output layer size differed between the methods. For IAIF-DNN, the two pulses were stretched to 400 samples, whereas only 300 samples were chosen for QCP-DNN (300 samples for a two-cycle segment corresponds to a f_0 of 106 Hz which was below the f_0 range of the female voice). As done previously in [13], initialization was performed without any pre-training, and the input vectors were scaled to lie between 0.1 and 0.9. Additionally for QCP-DNN, a Hann window was used for error weighting to emphasize the mid-signal excitation peak carrying important high-frequency components. Both networks were trained using the GPU-based Theano software [26, 27], which reduced the training time significantly compared to the previously used MATLAB-implementation.

An example of QCP-DNN generated glottal flow derivative waveforms is presented in Fig. 3. On top, 3(a) shows the DNN output when the input f_0 is varied while keeping the other input parameters constant. The variation can be seen to affect not only the generated pulse length, but also the sharpness of the main excitation peak in the middle. The corresponding two-pitch-cycle overlap-added waveforms are presented on bottom in 3(b) to better illustrate the effect of varying pitch in the synthetic excitation waveform.

3.3. Training of the HMM synthesis systems

The three synthesis systems were trained using the HTS 2.3–Beta[1] HMM-synthesis toolkit [28], with the modification of the STRAIGHT based demo to accommodate our feature vectors. All

[1] http://hts.sp.nitech.ac.jp/?Download (accessed Sept. 2015)

Table 1. Scale used in the subjective evaluation.

3	much more natural
2	somewhat more natural
1	slightly more natural
0	equally natural
-1	slightly less natural
-2	somewhat less natural
-3	much less natural

systems use the same speech waveform data and the context dependent phonetic labels provided in the ARCTIC database for training. From the perspective of the HMMs, there is no difference between IAIF baseline and IAIF-DNN because they share their acoustic parametrization and only differ in their vocoder excitation methods.

4. SUBJECTIVE LISTENING TESTS

Subjective evaluation of the three speech synthesis systems was carried out by a pair comparison test based on the Category Comparison Rating (CCR) [29] test, where the listeners were presented with synthetic sample pairs produced from the same linguistic information with the different systems under comparison. The listeners were asked to evaluate the naturalness of first sample compared to the second sample using the seven point Comparison Mean Opinion Score (CMOS) scale presented in Table 1. The listeners were able to listen each pair as many times as they wished and the order of the test cases was randomized separately for each listener.

The listening was conducted with the TestVox[2] online application in a controlled listening environment by native English speaking listeners with no reported hearing disorders. In order to make the listening task more convenient, the listening experiment was partitioned in two tasks, with the first task containing eight different sentences and the second task containing seven. Null-pairs were included in the test and each test case was presented twice to ensure listener consistency and enable the possible post-screening of test participants. 14 listeners participated in the first part and 13 in the second, with some overlap in the participants. For analysis, the results were pooled together and null-pairs were omitted. No listeners were excluded in the analysis of results.

The result of the subjective evaluation is presented in Fig. 4. The figure shows the mean score for each pair comparison in the CCR test on the horizontal axis with the 95% confidence intervals. In other words, Fig. 4 depicts the order of preference of the three synthesis methods by averaging for each method all the CCR scores the corresponding synthesizer was involved. For each comparison, the mean difference was found to differ from zero with $(p < 0.001)$, indicating statistically significant listener preferences between the three synthesis methods. Corroborating our previous findings reported in [14], the results show small, yet significant difference between IAIF baseline IAIF-DNN in favor of the latter. Most strikingly, the proposed QCP-DNN method achieves a clearly higher score compared to both IAIF baseline and IAIF-DNN.

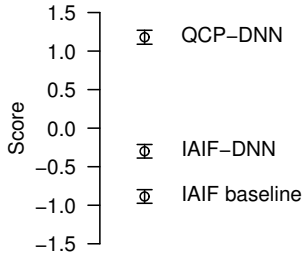


Fig. 4. Result of the subjective listening test on synthesized female voice naturalness.

5. CONCLUSIONS

Glottal vocoding aims to parameterize the speech signal by using a physiologically motivated approach by modelling the real excitation of the human voice production mechanism, the glottal flow. Given the recent success of this approach in, for example, synthesis of male voices [11] and in adaptation of the speaking style [15], the present study was launched to specifically focus on synthesis of female speech. Three glottal vocoders were compared: (1) the baseline glottal vocoder, GlottHMM introduced in [11], and (2) its recently developed new version based on DNNs [13], and (3) a new version that combines DNNs and a new glottal inverse filtering method, Quasi Closed Phase (QCP). In addition to the utilization of a new glottal inverse filtering method, the new vocoder, QCP-DNN, also introduces other modifications to its predecessors: the DNN excitation generation was modified so that glottal waveform is not interpolated in the training, leading to richer high frequency content in the generated excitation.

The three methods were trained to synthesise an US English female voice. Subjective evaluations were conducted with native English listeners using a CCR-type of evaluation. This evaluation showed that the proposed QCP-DNN method clearly outperforms the other two glottal vocoding methods, IAIF baseline and IAIF-DNN. This is likely due to, first, the more consistent vocal tract spectral representation given by QCP and second, a better quality of the inverse filtered glottal excitation used in the DNN training. The results are highly encouraging in showing that the subjective quality of synthesized female speech can be improved by utilizing new, more accurate physiologically motivated glottal vocoding techniques.

Future work includes incorporating the new vocoder to a DNN-based speech synthesis system, creating a more generalized speaker independent QCP-DNN based excitation method, and thorough subjective evaluation with a more extensive range of different speakers.

[2]https://bitbucket.org/happyalu/testvox/wiki/Home (accessed Sept. 2015)

6. REFERENCES

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

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

[3] Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura, "Speech synthesis based on hidden markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

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

[5] Zhen-Hua Ling, Shi-Yin Kang, Heiga Zen, Andrew Senior, Mike Schuster, Xiao-Jun Qian, Helen Meng, and Li 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] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain 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.

[7] Hideki Kawahara, Jo Estill, and Osamu 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.

[8] Harald Pobloth and W. Bastiaan Kleijn, "On phase perception in speech," in *Proc. of ICASSP*, Mar 1999, vol. 1, pp. 29–32 vol.1.

[9] Tuomo Raitio, Lauri Juvela, Antti Suni, Martti Vainio, and Paavo Alku, "Phase perception of the glottal excitation of vocoded speech," in *Proc. of Interspeech*, Dresden, September 2015, pp. 254–258.

[10] Tuomo Raitio, Antti Suni, Hannu Pulakka, Martti Vainio, and Paavo Alku, "HMM-based Finnish text-to-speech system utilizing glottal inverse filtering," in *Proc. of Interspeech*, Brisbane, Australia, September 2008, pp. 1881–1884.

[11] Tuomo Raitio, Antti Suni, Junichi Yamagishi, Hannu Pulakka, Jani Nurminen, Martti Vainio, and Paavo 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.

[12] Antti Suni, Tuomo Raitio, Martti Vainio, and Paavo Alku, "The GlottHMM speech synthesis entry for Blizzard Challenge 2010," in *Blizzard Challenge 2010 Workshop*, Kyoto, Japan, September 2010.

[13] Tuomo Raitio, Heng Lu, John Kane, Antti Suni, Martti Vainio, Simon King, and Paavo Alku, "Voice source modelling using deep neural networks for statistical parametric speech synthesis," in *22nd European Signal Processing Conference (EUSIPCO)*, Lisbon, Portugal, September 2014.

[14] Tuomo Raitio, Antti Suni, Lauri Juvela, Martti Vainio, and Paavo 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.

[15] Tuomo Raitio, Antti Suni, Martti Vainio, and Paavo Alku, "Synthesis and perception of breathy, normal, and Lombard speech in the presence of noise," *Computer Speech & Language*, vol. 28, no. 2, pp. 648–664, March 2014.

[16] Ling-Hui Chen, Tuomo Raitio, Cassia Valentini-Botinhao, Zhen-Hua Ling, and Junichi Yamagishi, "A deep generative architecture for postfiltering in statistical parametric speech synthesis," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 11, pp. 2003–2014, Nov 2015.

[17] Manu Airaksinen, Tuomo Raitio, Brad Story, and Paavo 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.

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

[19] Antti Suni, Tuomo Raitio, Martti Vainio, and Paavo Alku, "The GlottHMM entry for Blizzard Challenge 2011: Utilizing source unit selection in HMM-based speech synthesis for improved excitation generation," in *Blizzard Challenge 2011 Workshop*, Turin, Italy, September 2011.

[20] Paavo Alku, "Glottal inverse filtering analysis of human voice production – a review of estimation and parameterization methods of the glottal excitation and their applications. (invited article)," *Sadhana – Academy Proceedings in Engineering Sciences*, vol. 36, no. 5, pp. 623–650, 2011.

[21] John Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, Apr 1975.

[22] Paavo Alku, Jouni Pohjalainen, Martti Vainio, Anne-Maria Laukkanen, and Brad Story, "Formant frequency estimation of high-pitched vowels using weighted linear predictiona)," *The Journal of the Acoustical Society of America*, vol. 134, no. 2, 2013.

[23] Thomas Drugman and Thierry Dutoit, "Glottal closure and opening instant detection from speech signals.," in *Proc. of Interspeech*, 2009, pp. 2891–2894.

[24] Thomas Drugman, Mark Thomas, Jon Gudnason, Patrick Naylor, and Thierry Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 994–1006, March 2012.

[25] John Kominek and Alan W. Black, "CMU ARCTIC databases for speech synthesis," Tech. Rep., Language Technologies Institute.

[26] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio, "Theano: a CPU and GPU math expression compiler," in *Proc. of the Python for Scientific Computing Conference (SciPy)*, June 2010, Oral Presentation.

[27] Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, and Yoshua Bengio, "Theano: new features and speed improvements," Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

[28] Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W. Black, and Keiichi Tokuda, "The HMM-based speech synthesis system version 2.0," in *Proc. of ISCA SSW6*, Bonn, Germany, August 2007, pp. 294–299.

[29] "Methods for Subjective Determination of Transmission Quality," Recommendation P.800, ITU-T SG12, Geneva, Switzerland, Aug. 1996.