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

GlottDNN — A full-band glottal vocoder for statistical parametric speech synthesis

Manu Airaksinen[1], *Bajibabu Bollepalli*[1], *Lauri Juvela*[1], *Zhizheng Wu*[2], *Simon King*[2], *Paavo Alku*[1]

[1]Department of Signal Processing and Acoustics, Aalto University, Finland
[2]The Centre for Speech Technology Research, University of Edinburgh, United Kingdom

manu.airaksinen@aalto.fi, paavo.alku@aalto.fi

Abstract

GlottHMM is a previously developed vocoder that has been successfully used in HMM-based synthesis by parameterizing speech into two parts (glottal flow, vocal tract) according to the functioning of the real human voice production mechanism. In this study, a new glottal vocoding method, GlottDNN, is proposed. The GlottDNN vocoder is built on the principles of its predecessor, GlottHMM, but the new vocoder introduces three main improvements: GlottDNN (1) takes advantage of a new, more accurate glottal inverse filtering method, (2) uses a new method of deep neural network (DNN) -based glottal excitation generation, and (3) proposes a new approach of band-wise processing of full-band speech.

The proposed GlottDNN vocoder was evaluated as part of a full-band state-of-the-art DNN-based text-to-speech (TTS) synthesis system, and compared against the release version of the original GlottHMM vocoder, and the well-known STRAIGHT vocoder. The results of the subjective listening test indicate that GlottDNN improves the TTS quality over the compared methods.

Index Terms: speech synthesis, vocoder, glottal inverse filtering, deep neural network

1. Introduction

Statistical parametric speech synthesis [1], along with unit selection synthesis [2], is one of the two main disciplines in text-to-speech (TTS) synthesis. Whereas research of unit selection synthesis has reached a mature status in the past few years, the technology in parametric speech synthesis is currently under extensive progress. In particular, systems with hidden Markov model (HMM) -based back-ends have been replaced increasingly with deep neural network (DNN) -based technologies [3], and more recently, with long-short-term memory (LSTM) [4] -based systems with significant improvements in quality. Statistical TTS, however, suffers from two drawbacks: "muffledness" and "buzziness" [1]. Muffledness is caused by over-smoothing of parameter trajectories due to averaging by HMMs, whereas buzziness is caused by using overly simplified, impulse-like excitation waveforms in vocoding of voiced speech. While advent of DNNs and LSTMs has helped in tackling muffledness, buzziness still severely degrades the quality and naturalness of current statistical TTS systems.

The vocoders used in statistical parametric speech synthesis can be divided into three main categories: Mixed/impulse-excited vocoders (e.g. STRAIGHT [5, 6]), glottal vocoders (e.g. GlottHMM [7]), and sinusoidal vocoders (e.g. Quasi-harmonic model [8]). The first two categories utilize the source-filter model of speech production [9] which assumes that speech is produced by a source signal that is convolved with a filter conveying the vocal tract formants. The difference between the mixed/impulse-excited and glottal vocoders is in the interpretation of the vocoder excitation: The mixed excitation approach assumes that the excitation signal is spectrally flat and contains the pitch, noise, and phase information, and the filter models the entire spectral envelope of the signal. The glottal vocoding approach in turn assumes a more physiologically motivated distinction between the excitation and the filter: The excitation is assumed to correspond to the time-derivative of the true airflow generated at the vocal folds (consisting of the combined effects of the glottal volume velocity and lip radiation [9]), and the filter corresponds to a transfer function that is created by the physiological organs of the human vocal tract. Since natural talkers are capable of varying the vibration mode of their vocal folds, the spectral envelope of the glottal excitation is, importantly, not constant (e.g. flat) but instead varies when, for example, the phonation type of speech changes [10].

Since its inception, the GlottHMM vocoder proposed in Raitio et al. [7] has become a relatively well-known glottal vocoding platform [1] that received its final release version recently. During its development phase, the vocoder underwent multiple changes in its components (e.g. changes in glottal pulse selection and generation [11], glottal inverse filtering method [12]). Our previous studies with GlottHMM, conducted almost exclusively with 16 kHz speech, have been successful particularly in high-quality synthesis of a Finnish male voice [7], in adapting the synthesis to different speaking styles [13] and in improving the synthesis intelligibility under noisy conditions [14]. Recently, we introduced a novel DNN-based glottal excitation generation scheme aiming specifically at high-pitched voices [15]. In the proposed excitation generation, a DNN is trained to generate a glottal flow derivative waveform computed with a new glottal inverse filtering method, quasi-closed phase analysis (QCP) [16]. The proposed vocoder was evaluated in [15] with a HMM-based synthesis system resulting in significantly improved performance over two baseline glottal vocoders.

Given the high potential of glottal vocoding indicated both by GlottHMM and the more recent QCP-based technique, the current study examines, for the first time, the use of QCP-based glottal vocoding in synthesis of full-band (48 kHz) speech. The study introduces a new glottal vocoder, named GlottDNN. The new vocoder involves many improvements to its predecessors, both GlottHMM and the QCP-based method introduced in [15] (e.g. new method of DNN-based glottal pulse generation, new approach of band-wise processing of speech), all of which are described in detail in Section 2. The proposed new GlottDNN vocoder is used in a state-of-the-art DNN-based synthesis plat-

http://dx.doi.org/10.21437/Interspeech.2016-342

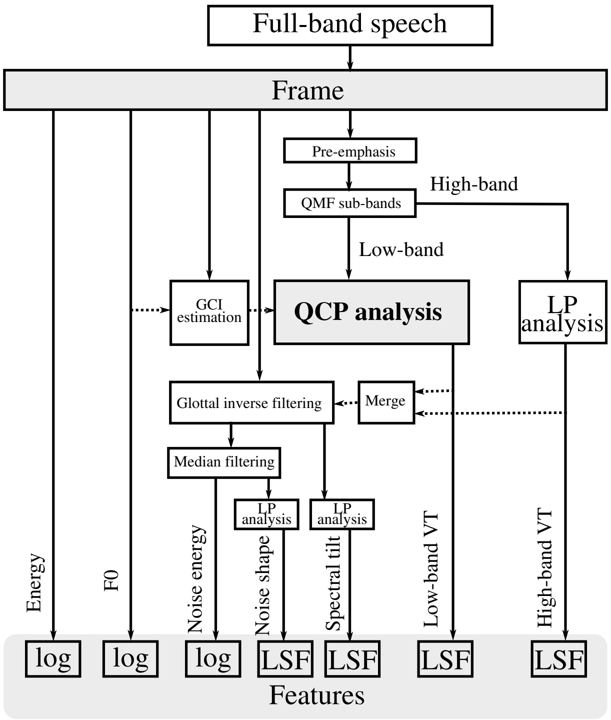

Figure 1: *Block diagram of the analysis stage.*

form proposed in [17] to synthesize a 48 kHz male voice and the vocoder performance is compared to the release versions of GlottHMM and STRAIGHT in a subjective listening test.

2. The GlottDNN vocoder

2.1. General

The main changes from the most widely used previous glottal vocoder, GlottHMM, to the new GlottDNN vocoder can be categorized as follows: (1) The GlottDNN vocoder utilizes a recently proposed glottal inverse filtering (GIF) method, quasi-closed phase (QCP) analysis. QCP [16] enables more accurate estimation of the glottal flow than iterative adaptive inverse filtering (IAIF) [18], the inverse filtering method used in GlottHMM. (2) The glottal excitation of the synthesized speech is generated with a specific DNN that maps vocoder feature vectors into time-domain glottal pulses. [15] (3) GlottDNN supports full audio band (up to 48 kHz sampling rate) that is an increasingly imposed demand for state-of-the-art quality.

2.2. Full-band speech analysis and parametrization

The block diagram depicting the GlottDNN's analysis stage for voiced frames is shown in Figure 1, and the extracted parameters are presented in Table 1. First, frame energy and fundamental frequency (f_0) information is extracted and saved to the feature vectors. Next, the pre-emphasized frame is split into low- and high-band signals with quadrature mirror filtering (QMF) [19] using a linear phase a FIR filter with a cut-off frequency of $0.5 \cdot \pi$. Then, the bands are down-sampled by a factor of two. For full-band speech sampled with 48 kHz, the low-band and high-band cover the frequency range of 0 Hz – 12 kHz and 12 kHz – 24 kHz, respectively. Next, the analysis is split between the high-band and low-band.

Table 1: *Speech features and the number of parameters used for full-band (48 kHz) speech. The parameter orders were determined heuristically based on informal listening tests.*

Feature	Parameters per frame
Fundamental frequency (log F0)	1
Energy (log)	1
Low-band vocal tract (LSF)	42
High-band vocal tract (LSF)	18
Spectral tilt (LSF)	24
Noise shape (LSF)	24
Noise energy (log)	1

QMF-based sub-bands are used in GlottDNN because of the following two drawbacks in auto-regressive (AR) modeling of wide-band speech. (1) Line spectral frequencies (LSFs) are widely used to represent AR filters and they have been taken advantage of, for example, in previous glottal vocoders [7]. Using classical root solving techniques (e.g. [20]), however, to convert LSFs to polynomials and vice versa results in severe accuracy problems with large sampling frequencies, such as 48 kHz, because AR-model orders become inevitably large when set according to the rule [9] between the sampling frequency (F_s) and the filter order (p) (e.g. $p > 50$ with $F_s = 48$ kHz). (2) The perceptually most important part of the vocal tract is the frequency range of 0.5 kHz– 4 kHz containing the first three formants. Improved spectral modeling of this frequency range in full-band speech would in principle be possible by using frequency warping [21] in the computation of AR models without the need for using large AR orders. The use of frequency warping in the GlottDNN vocoder is not justified because the vocoder is based on glottal inverse filtering (i.e. QCP), a procedure which cancels the vocal tract resonances on a linear frequency scale. By splitting the speech signal with the proposed QMF approach into the low-band and high-band, it is, however, possible to utilize AR models whose order is smaller due to narrower width of the two bands. In addition, the procedure enables allocating more accurate, larger order AR models for the more important low-band while using smaller order spectral models for the high-band.

The low-band contains the most important characteristics of the glottal excitation, including the glottal formant as well as the most prominent harmonics [9]. To decouple between the glottal excitation and the vocal tract formant structure, a high-quality GIF method, quasi-closed phase (QCP) analysis [16], is applied to the low-band signal hence obtaining an estimate of the vocal tract (VT) filter envelope. QCP is based on weighted linear prediction (WLP) [22] in which the square of the prediction error is temporally weighted with a pre-defined weighting function. By using weighting functions that attenuate the prediction error near glottal closure instants (GCIs), WLP enables computing vocal tract models that are less biased by the glottal source. Since the glottal excitation has a reduced effect within the high-band (e.g. harmonics are less prominent), it is possible to model the high-band envelope with conventional LP analysis. The low- and high-band VT estimates are converted into LSFs saved to the feature vector.

Next, the transfer functions of the low- and high-band filters are combined into a single polynomial to be used in glottal inverse filtering of the frame. The zero-padded discrete Fourier transforms (DFT) of length N_{FFT} are computed for the two filter polynomials, and their magnitude spectra are concatenated to form a single magnitude spectrum of length $2N_{\text{FFT}}$. The gain

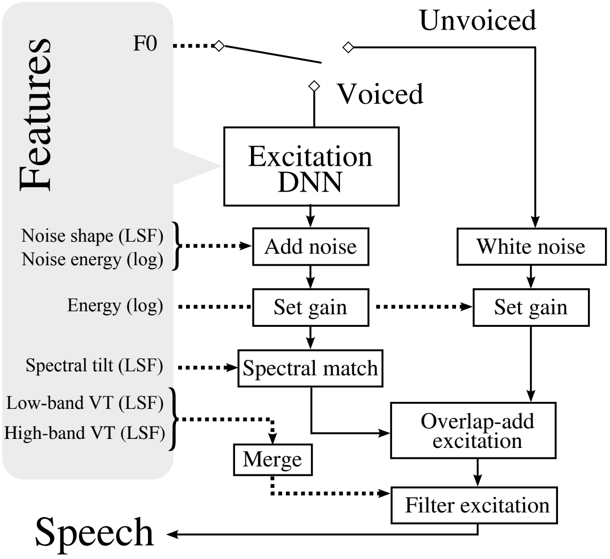

Figure 2: *Block diagram of the synthesis stage.*

of the high-band spectrum is set so that the amplitude at its first sample matches that of the last sample of the low-band spectrum. Then, the square of the concatenated magnitude spectrum is inverted, and the result is inverse Fourier transformed (IDFT) to obtain the corresponding autocorrelation sequence. The autocorrelation function is finally transformed with the Levinson-Durbin recursion [9] to obtain a stable all-pole filter modeling the vocal tract of the full-band.

The obtained full-band vocal tract model is used for inverse filtering the speech signal. The noise component of the glottal flow derivative is estimated with median filtering. The median filter uses a window size of 4 ms and the filtering is computed in two parts for a two pitch-period glottal flow derivative waveform using square-root Hann windowing so that the main excitation peak in the center is kept intact. By subtracting the median filter output from the glottal flow derivative, a noise-like signal, called median filter residual, is obtained. The median filter residual is considered to represent the desired noise component of the excitation and it is further high-pass filtered with a cut-off frequency of 2 kHz. The energy of the median filter residual relative to the original glottal flow derivative is then parametrized into the feature vector. Finally, the spectral shape of the median filter residual is parametrized with LP analysis. The main justifications in using the proposed approach in estimating the noise component from voiced speech are: the process is (1) robust and (2) computationally effective. The corresponding harmonic-to-noise ratio (HNR) -based modeling of the noise component in GlottHMM requires iterative manipulation in the spectral domain, which is computationally heavy, especially for full-band speech where the vector sizes are large.

For unvoiced frames, the analysis procedure is greatly simplified. Both the low- and high-band VT components are modeled with LP analysis, and along with the energy, they are solely used to model the frame.

2.3. Full-band speech synthesis

Speech synthesis from the vocoder parameters of Table 1 is depicted in the block diagram of Figure 2. First, the entire excitation signal is generated separately for voiced and unvoiced frames. The voiced excitation signal is generated with pitch-synchronous overlap-add (PSOLA) of two pitch-period glottal flow derivative pulses. The pulses are generated with a deep neural network (DNN) (more details in [15]) that predicts a two pitch-period glottal flow derivative wave, estimated in the training phase with QCP, from an acoustic feature vector. The pulse waveform is square root Hann windowed and aligned to have a glottal closure instant at the center of a fixed-length frame. The signal is zero-padded from its edges to keep its original length unchanged. After the pulse generation, the signal is windowed with a square root Hann window whose length is equal to twice the length of the desired fundamental period. Next, the (voiced speech) noise component is generated as uniformly distributed white noise whose spectral envelope and intensity is shaped by the corresponding LSFs and energy value. Finally, the processed signal is added to the DNN-generated excitation waveform.

Before PSOLA, the generated glottal excitation is matched in terms of its spectral tilt to the target pulse. This is done by using a pole-zero matching filter that is constructed as a ratio between two low-order LP filters of an equal prediction order:

$$H_{\text{match}}(z) = \frac{H_{\text{base}}(z)}{H_{\text{target}}(z)}, \tag{1}$$

where $H_{\text{match}}(z)$ is the matching filter, and $H_{\text{base}}(z)$ denotes the LP inverse model of the frame to be matched and $\frac{1}{H_{\text{target}}}(z)$ is the target spectral tilt.

Even though the above described process of generating the glottal excitation in the GlottDNN vocoder might look similar to that of GlottHMM, the two vocoders differ greatly in excitation generation. In GlottDNN, the DNN-based excitation generation utilizes the QCP inverse filtering algorithm as opposed to the previously used IAIF-based method [13], and produces glottal flow derivative pulses that do not need to be interpolated to match a target length. This solves many problems: (1) The obtained base pulse has more realistically varying gross shape from frame to frame when compared to the original GlottHMM where either a single base pulse is used [7] or the best fitting pulse is searched from a library of base pulses [11]. (2) The obtained pulse requires less processing than in GlottHMM, as its length does not need to be interpolated.

The unvoiced excitation is generated frame-by-frame as spectrally white noise whose energy is scaled to match the target. After the voiced and unvoiced excitation waveforms have been generated, they are combined, and the obtained excitation is filtered with the vocal tract model obtained from the low-band and high-band LSFs. The low-band and high-band filters are merged with the same process as described in Section 2.2.

3. Experiments

The GlottDNN vocoder was evaluated with a subjective comparison category rating (CCR) listening test [23] on the naturalness of TTS quality. The comparison was conducted between GlottDNN, the baseline GlottHMM vocoder (using frequency warping as proposed in [21]), and the STRAIGHT vocoder [5, 6] that has become the de-facto standard in statistical parametric speech synthesis.

Speech data employed in our experiments consisted of 2572 utterances recorded by a male British speaker named as “Nick” [24]. Those utterances were divided into 2400, 70, and 72 utterances as training set, development set and evaluation set, respectively. The sampling rate of the speech data was 48 kHz.

Table 2: *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

3.1. The TTS system

The listening test samples were created with a state-of-the-art DNN-based TTS system proposed in [17]. The TTS DNN system has 5 hidden layers with 1024 nodes in each layer. The activation function of hidden layers was tanh (hyperbolic tangent), and linear function for output layer. During training, weights were regularized by L2 norm with penalty factor of 0.00001, the mini-batch size was set to 256 and momentum was used. The maximum number of epochs was set to 25 with early stopping criteria.

The input features for all the three systems were equal, and extracted from the question file used for the decision tree clustering in the HTS system. The dimension was 601 which comprised 592 binary features and 9 numerical features. The binary features contains the information about such as quinphone identity, syllable location, part-of-speech, word and phrase. The appended numerical features provide the information at frame level such as the frame position within the HMM state and phoneme, the state position within the phoneme, and state and phoneme durations. Min-max normalization was applied on input features which scaled the features into the range of [0.1, 0.99]. Mean-variance normalization was applied on the output features. MLPG using pre-computed variances from the training data was applied to the output features.

3.2. Subjective evaluation

Subjective evaluation of the three speech synthesis systems was carried out by a pair comparison test based on the Category Comparison Rating (CCR) 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 2. 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.

A web based listening test was conducted with the modified Beaqlejs application [25]. 10 synthesized samples were selected from each system and 10 null-pairs were included in the test. Each test case was presented twice to ensure listener consistency and enable the possible post-screening of test participants, which results in a total of 60 (3*2*10) + 10 = 70 samples for the listening test. In order to reduce the duration of the listening test, we presented only 35 samples, selected randomly from the 70 samples, for each subject.

A total of 12 subjects, mainly international master's students at Aalto University and University of Edinburgh, participated in the listening test. All subjects were included in the final analysis of the results. The results of the subjective evaluation are presented in Figure 3. 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, Figure 3 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. The results indicate that the glottal vocoders, GlottHMM and GlottDNN, provide superior quality over the STRAIGHT system, and the proposed GlottDNN system improves the quality over the baseline GlottHMM system. All of these differences have a statistically significant margin.

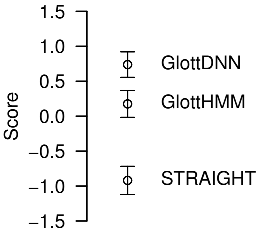

Figure 3: *Results of the subjective listening test with their 95% confidence intervals.*

4. Discussion

A new glottal vocoding method, GlottDNN, was proposed in the present study. GlottDNN utilizes a method of glottal vocoding where a new glottal inverse filtering method, quasi-closed phase analysis (QCP), is applied to full-band speech signals. In synthesis, a deep neural network (DNN) is used to predict a glottal excitation waveform from the vocoder parameters to obtain more natural synthesis quality.

GlottDNN was evaluated in synthesis of a full-band (48 kHz) male voice using a state-of-the-art DNN-based TTS system. The evaluation compared GlottDNN with the baseline GlottHHM vocoder and the well-known STRAIGHT vocoder in a comparison category rating (CCR) test. The results of the CCR test indicate that GlottDNN improves the synthesis naturalness with a statistically significant margin in relation to the compared methods.

Plans of future research involving the GlottDNN vocoder involve a more thorough investigation on the vocal tract spectrum modeling representation, and the evaluation of performance on a more broad set of speakers and speaking styles.

5. Acknowledgements

The research leading to these results has received funding from the Academy of Finland (project no. 256961, 284671).

6. References

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

[2] A. W. Black and N. Campbell, “Optimising selection of units from speech databases for concatenative synthesis,” in *In Proc. Eurospeech*, 1995, pp. 581–584.

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

[4] H. Zen and H. Sak, “Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis,” in *Proc. ICASSP*, 2015, pp. 4470–4474.

[5] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, “Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based {F0} extraction: Possible role of a repetitive structure in sounds1,” *Speech Communication*, vol. 27, no. 34, pp. 187 – 207, 1999.

[6] 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 *Proc. MAVEBA*, 2001.

[7] 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, 2011.

[8] 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, 2014.

[9] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*, ser. Prentice-Hall signal processing series. Prentice-Hall, 1978.

[10] S. A. Zollinger and H. Brumm, “The lombard effect,” *Current Biology*, vol. 21, no. 16, pp. R614 – R615, 2011.

[11] 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 *Proc. ICASSP*, 2011.

[12] A. Suni, T. Raitio, M. Vainio, and P. 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*, 2011.

[13] 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. Interspeech*, 2014.

[14] A. Suni, T. Raitio, M. Vainio, and P. Alku, “The glotthmm speech synthesis entry for blizzard challenge 2010,” in *Blizzard Challenge 2010 Workshop*, 2010.

[15] 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. ICASSP*, 2016.

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

[17] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, “Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis,” in *Proc. ICASSP*. IEEE, 2015, pp. 4460–4464.

[18] P. Alku, “Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering,” *Speech Communication*, vol. 11, no. 2–3, pp. 109 – 118, 1992.

[19] J. Johnston, “A filter family designed for use in quadrature mirror filter banks,” in *Proc. ICASSP*, vol. 5, 1980, pp. 291–294.

[20] A. Edelman and H. Murakami, “Polynomial roots from companion matrix eigenvalues,” *Math. Comp*, vol. 64, pp. 763–776, 1995.

[21] T. Raitio, A. Suni, M. Vainio, and P. Alku, “Wideband parametric speech synthesis using warped linear prediction,” in *Proc. Interspeech*, 2012.

[22] C. Ma, Y. Kamp, and L. Willems, “Robust signal selection for linear prediction analysis of voiced speech,” *Speech Communication*, vol. 12, no. 1, pp. 69 – 81, 1993.

[23] ITU, “Methods for subjective determination of transmission quality,” in *International Telecommunication Union, Recommendation ITU-T P.800*, 1996.

[24] M. Cooke, C. Mayo, and C. Valentini-Botinhao, “Hurricane natural speech corpus,” 2013, LISTA Consortium. [Online]. Available: http://dx.doi.org/10.7488/ds/140

[25] S. Kraft and U. Zölzer, “BeaqleJS: HTML5 and JavaScript based Framework for the Subjective Evaluation of Audio Quality,” in *Linux Audio Conference*, 2014.