Airaksinen, Manu; Alku, Paavo

Effects of training data variety in generating glottal pulses from acoustic features with DNNs

Published in:
Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH

DOI:
10.21437/Interspeech.2017-363

Published: 01/08/2017

Document Version
Publisher's PDF, also known as Version of record

INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Effects of training data variety in generating glottal pulses from acoustic features with DNNs

Manu Airaksinen, Paavo Alku

Aalto University, Finland
manu.airaksinen@aalto.fi, paavo.alku@aalto.fi

Abstract

Glottal volume velocity waveform, the acoustical excitation of voiced speech, cannot be acquired through direct measurements in normal production of continuous speech. Glottal inverse filtering (GIF), however, can be used to estimate the glottal flow from recorded speech signals. Unfortunately, the usefulness of GIF algorithms is limited since they are sensitive to noise and call for high-quality recordings. Recently, efforts have been taken to expand the use of GIF by training deep neural networks (DNNs) to learn a statistical mapping between frame-level acoustic features and glottal pulses estimated by GIF. This framework has been successfully utilized in statistical speech synthesis in the form of the GlottDNN vocoder which uses a DNN to generate glottal pulses to be used as the synthesizer's excitation waveform. In this study, we investigate how the DNN-based generation of glottal pulses is affected by training data variety. The evaluation is done using both objective measures as well as subjective listening tests of synthetic speech. The results suggest that the performance of the glottal pulse generation with DNNs is affected particularly by how well the training corpus suits GIF: processing low-pitched male speech and sustained phonations shows better performance than processing high-pitched female voices or continuous speech.
Index Terms: glottal inverse filtering, speech synthesis

1. Introduction

The source signal of voiced speech, the glottal volume velocity waveform (also known as the glottal flow), is responsible for generating some of the most essential acoustical cues in speech such as pitch and the type of phonation. In addition, the glottal flow carries information, for example, about the emotional state of the speaker, individual speech characteristics, and possible voice pathologies. Within the human speech production mechanism, the glottal flow is modulated by the resonances of the vocal tract that convey linguistic information. This acoustic linkage of the glottal flow and the vocal tract hinders obtaining direct information about the glottal flow, as direct measurements at the glottis in conjunction with natural speech production are not possible. Glottal inverse filtering (GIF) aims to non-invasively estimate the glottal flow from a recorded microphone speech signal by applying such anti-resonances to the speech signal that cancel the effects of the vocal tract [1].

The utility domain of GIF is, unfortunately, limited, as GIF algorithms are sensitive to noise and call for high-quality recordings [1], which mostly make the use of GIF infeasible outside laboratory conditions. Recently, efforts have been taken to expand the use of GIF by training neural networks (e.g., deep neural networks (DNNs) [2] or long short-term memory networks (LSTMs) [3]) to learn a statistical mapping between frame-level acoustic features and glottal flows estimated by GIF. So far, these networks have mainly been used in statistical speech synthesis within the context of a glottal vocoder [4, 5] where the vocoder parameters are used to generate the underlying glottal flow waveform.

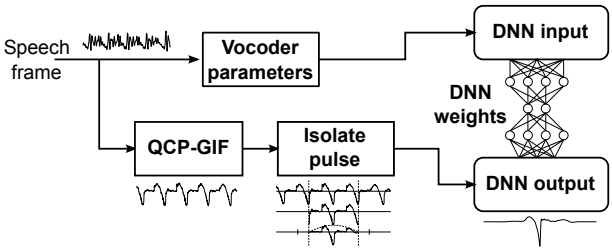

Figure 1: *Training procedure for the deep neural network-based glottal pulse generation.*

In order to learn a statistical mapping between acoustic features and the corresponding glottal flow, a question about the effects of the training data arises. It is not clear, for instance, whether using a voice specific training as opposite to a multi-speaker training makes a difference in the deep learning -based generation of the glottal flow. Also, the performance of signal processing algorithms (i.e., GIF, glottal closure instant detection) might constrain the variety of the training data.

The goal of this investigation is to study, using both objective and subjective performance measures, how variety of the training data affects the deep learning -based glottal pulse generation. The following section describes first in detail the glottal pulse generation framework. Then the extraction process to collect the training data and its most common error sources are addressed in Section 3. Finally, experiments on the effects of training data variety are reported and analyzed in Sections 4 and 5.

2. Neural network-based glottal pulse generation

The neural network-based glottal pulse generation is based on defining a statistical mapping between a compressed set of frame-level speech features (such as F0 and spectral envelope) and the corresponding glottal flow waveform. Despite constantly improving end-to-end machine learning paradigms (e.g., WaveNet [6]), the task of glottal source estimation, however, cannot be out-sourced entirely to machine learning, as the target glottal flow waveforms that would act as reference in the network training cannot be directly recorded from natural speech [7, 1]. The training of the network thus heavily relies on the estimation of the reference glottal pulses which can be conducted only via GIF. The actual task of the glottal pulse-generating network in turn is just to map frame-level features into outputs computed by GIF. The neural network-based glottal flow

http://dx.doi.org/10.21437/Interspeech.2017-363

generation can be used, for example, when speech recordings are of diminished quality and conventional GIF analysis fails to perform correctly [8]. However, the focus of the current study is on the DNN-based glottal pulse generation within statistical parametric speech synthesis (SPSS).

Using vocoding with over-simplified excitation models is known to be one of the main reasons for quality degradation in SPSS [9]. In glottal vocoders, such as the GlottHMM vocoder [10], improved quality is achieved by substituting the traditional spectrally flat impulse-plus-noise excitation [11, 12] with a glottal flow waveform that is generated using a unit selection approach from a library of pre-computed glottal pulse waveforms. This approach, however, is problematic with respect to the pulse selection. In the original GlottHMM vocoder, a single base pulse was used to generate the entire glottal excitation waveform. In later experiments, larger pulse libraries were implemented, but they failed to provide significant improvements over the baseline [13]. This is both due to having low-quality glottal waveforms in the pulse library, caused by poor GIF performance, and due to introducing ambiguity in the objective criteria of the target cost: The glottal pulses were parameterized according to their spectral features and F_0, but the pulses were interpolated to constant length. The interpolation introduces problems by shifting the glottal formant [14], and the selection of the best objective criterion for pulse search within a large database of candidates is not clear. The DNN-based excitation generation was introduced in [2] and [15] to overcome the problems described above: First, the neural network learns to map the input features to the desired output automatically, so heuristic, hand-crafted objective criteria can be avoided. Second, as the neural network can be trained on large amounts of data, and the outputs reflect the statistical likelihoods of the training data conditional to the input, outlier pulses will not be generated.

The original DNN-based glottal pulse framework presented in [2, 15] was recently successfully modified in [4] where the Iterative Adaptive Inverse Filtering (IAIF) [16] GIF method was replaced with a new, more accurate Quasi-closed Phase (QCP) [17] algorithm, and the pulse interpolation was changed to a combination of half-sine windowing and zero-padding. This framework, presented in Figure 1, is assumed in the rest of the paper. In the framework, a deep neural network, more precisely a multi-layer perceptron with three hidden layers, is trained to learn the mapping between the input, a set of the GlottDNN vocoder parameters, and the output, an isolated glottal flow derivative waveform computed by the QCP algorithm.

3. Training data extraction with glottal inverse filtering

3.1. Glottal pulse extraction procedure

The training data extraction process for the DNN-based glottal pulse generation is illustrated in Figure 2(a). First, a glottal flow signal is estimated from the input, a voiced speech frame, with GIF and differentiated. (From now on, the differentiated glottal flow is called the voice source). This voice source estimate can be computed with different analysis parameters (e.g., analysis frame length, vocal tract filter order) from the input acoustical parameters. Next, a two-pitch-period segment is selected from the voice source, delimited by glottal closure instants (GCIs) that have been computed with a specific GCI estimation algorithm (e.g., [18]). Finally, the two-pitch-period segment is half-sine windowed, and zero-padded from both sides to a constant length. The windowing function must be compatible with the overlap-add procedure (i.e. the same window is used once again after waveform generation) [19]. In this study, the Hann windowing was used but it is also possible to take advantage of more sophisticated overlap-add windowing functions, such as the Kaiser-Bessel derived window function [20]. The zero-padding is implemented by placing the negative minimum peak (approximately in the middle GCI) of the windowed pulse to the middle of the fixed-length vector. This procedure positions the most important temporal region of the voice source, the so-called "moment of maximum excitation" [21], to appear in a fixed position.

3.2. Error sources on training data extraction

For a given segment of speech, the waveform of the glottal flow estimate depends both on the inherent properties of the GIF algorithm used (e.g. IAIF vs. QCP) or the selected algorithm's hyper-parameters. The hyper-parameters of the QCP algorithm are the frame length w_l, vocal tract filter order p, GCI estimate vector $\mathbf{g} = [t_1, t_2, \ldots, t_n]$ where t_n denotes the nth GCI within the frame, and weighting function -related parameters (position quotient P, duration quotient D, and ramp size N_r) [17]. Since fixed values can be used for the weighting function -related parameters [17], the main hyper-parameters of interest become w_l, p, and $\mathbf{g}$.

3.2.1. Effect of short-time windowing

Speech tempo affects GIF analysis: Fast continuous speech is more problematic compared to slow continuous speech or sustained utterances, because the time-window over which the speech signal can be assumed to be stationary becomes shorter. Since GIF algorithms such as QCP operate under the assumption of stationarity, fast speech should in principle be analyzed with a smaller w_l value. With shorter analysis frames, however, the accuracy of GIF might deteriorate due to having a smaller number of data samples to define the vocal tract model. Moreover, fast speech is typically combined with high pitch, which is known to be a factor that reduces the accuracy of GIF [17]. As a practical example, it is considerably more difficult to compute accurate glottal flow estimates with GIF from continuous, expressive female speech than from slow male speech of neutral speaking style [22].

Effects caused by varying the frame length w_l are demonstrated in Figure 2. Two QCP analyses were conducted using the same set of hyper-parameters except for w_l which was either 50 ms (i.e. applicable for sustained vowels, shown in Figure 2(a)) or 25 ms (i.e. applicable for fast continuous speech, shown in Figure 2(b)). We can see that the glottal flow estimate in Figure 2(b) is distorted by impulse-like artifacts and formant ripple. If this kind of degradation is systematic, it will affect the DNN-based computation of glottal pulses introduced in Section 2 because the training data will bias the underlying DNN to learn distorted glottal flow waveforms.

3.2.2. Glottal closure instant estimation errors

If glottal pulses are estimated by GIF algorithms (such as QCP or traditional closed phase covariance analysis [23]) that call for defining GCIs, one of the main issues affecting the training data extraction process is the accuracy of the GCI estimation. First, in order to compute the vocal tract filter, these GIF algorithms require explicit GCI estimates. Inaccurate GCI estimates can lead to degraded glottal flow estimates akin to those shown in Figure 2(b). More importantly, even though the glot-

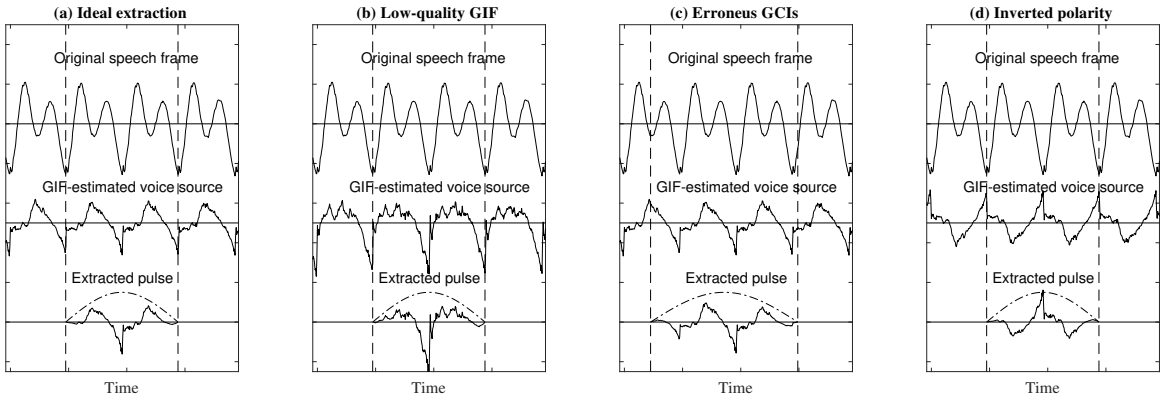

Figure 2: *Training data extraction process (a) with common error sources (b)-(d).*

tal flow waveform computed by GIF was free from distortion, the accuracy of GCIs is crucial to obtain training data that is correctly aligned, as illustrated in Figure 2(c). A misplaced or missed GCI can significantly shift the underlying phase of the pulse (which is assumed to be fixed), and the overall effect corresponds to having noisy training data as the fixed phase assumption has been violated.

3.2.3. Effect of the vocal tract filter order

Finally, the vocal tract filter order p can have multiple effects. In principle, one pair of complex conjugate poles in the z-domain can model one formant [24], so the ideal order of p depends on the sampling frequency f_s and the spectral complexity of the speech signal. If p is too low (meaning that the order of p is insufficient to account for all the vocal tract resonances within the signal's audio band), the resonances are not properly cancelled and the glottal flow is distorted by the so-called formant ripple [1] (see Figure 2(b)). On the other hand, if the order of p is much larger than the number of resonances, two error sources arise. First, the vocal tract filter of the GIF analysis starts to model the harmonic peaks of the excitation signal, which degrades source-filter separation. Second, tightly packed resonances might get too close to the zero-frequency, which can lead to polarity inversion illustrated in Figure 2(d).

4. Experiments

4.1. Speech databases

In order to understand how training data variety affects the generation of glottal pulses with DNNs, several speech corpora were used. The selected corpora are presented briefly in the following sections.

4.1.1. Blizzard challenge 2012 female voice "Nancy"

The "Nancy" corpus of Blizzard challenge 2012 [25] is a large, high-quality database including speech of a single female talker. This corpus is widely used in the text-to-speech synthesis community. The voice has been recorded by a professional American English voice actor, and it can be described as being on the expressive side of normal continuous speech. From the perspective of GIF, the "Nancy" voice is challenging, because it has f_0 contours that are both high-valued and rapidly changing.

4.1.2. Hurricane natural speech corpus male voice "Nick"

The "Nick" voice of the Hurricane natural speech corpus [26] is also a high-quality continuous speech corpus that is used in the TTS community. The voice is a British English male producing continuous speech that can be described as normal stable speech. From the perspective of GIF, the "Nick" voice is of the following characteristics: f_0 is low and the voice is of limited expressiveness and of high recording quality. Therefore, the "Nick" voice provides almost ideal conditions for the glottal pulse extraction in the context of continuous speech.

4.1.3. Voice conversion challenge 2016 database

The Voice Conversion Challenge (VCC) 2016 database [27] is a multi-speaker corpus consisting of 5 male and 5 female speakers. In this corpus, each speaker utters the same sentence set. The voices have been selected to be distinct from each other, which makes this corpus, among the databases of the current study, to be of the largest dynamics in glottal flow characteristics. This variety makes the VCC database challenging for the generation of glottal pulses, as the corpus contains several different and rapidly-changing voices.

4.1.4. Finnish sustained vowel database

As the last corpus, we selected a database of Finnish sustained vowel recordings. The database consists of three female and four male speakers all producing repeated sustained phonations of eight Finnish vowels ([a], [e], [i], [o], [u], [y], [ae], and [oe]) using a breathy, modal or pressed phonation type. The sustained vowel database is an ideal corpus to compute high-quality glottal pulses with GIF: Due to sustained phonation, the signals are stationary, which enables using longer analysis window lengths thus achieving better GIF quality (see Section 3.2.1). However, the acoustic space spanned by the sustained Finnish vowels might not be wide enough to cover the glottal source dynamics present in continuous English speech. The inclusion of multiple modes of phonation aims to reduce this problem by expanding the possible acoustic space. From the perspective of the DNN-based glottal pulse generation, a possible distinction between the sustained vowel database and the continuous speech databases might also be the distribution of glottal pulses: In the sustained vowel database, the number of glottal pulses is balanced between the three different phonation types, whereas in the continuous speech databases, one type of phonation (modal) dominates.

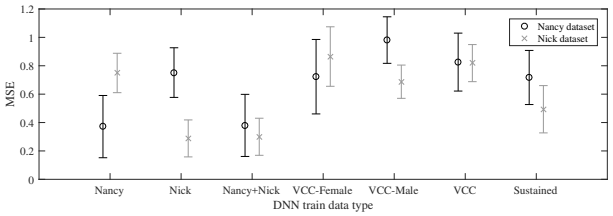

Figure 3: *Objective error scores.*

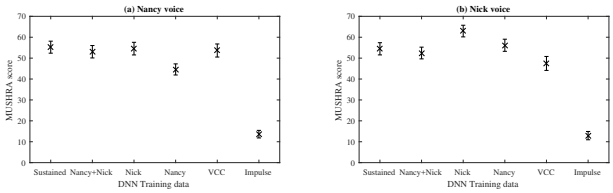

Figure 4: *Subjective MUSHRA listening test results.*

4.2. Glottal pulse generation DNN systems

The speech corpora presented in Section 4.1 were used to train 7 distinct DNNs to generate glottal pulses: "Nancy", "Nick", "Nancy+Nick", "VCC", "VCC-Male", "VCC-Female", and "Sustained". With this procedure, we aimed to obtain varying systems of interest. The "Nancy" DNN and the "Nick" DNN both correspond to voice-specific training using high-quality continuous speech recordings. The "Nancy+Nick" DNN is trained with a combination of a male voice and a female voice. The VCC corpus is used to build three DNNs: One trained with both genders (the VCC DNN), and two trained in a gender-specific manner (the VCC-Female DNN using female voices, the VCC-Male DNN using male voices). Finally, the sustained vowel database was used to train the Sustained DNN.

The same training data extraction procedure was used for the continuous speech databases: The GlottDNN vocoder [5, 3] was used to extract glottal pulses and parameterize the input using a 25-ms frame length and a 5-ms frame skip with a 16-kHz sampling rate. The vocoder parameters include f_0, energy, harmonic-to-noise ratio, vocal tract transfer function LSFs ($p = 30$), and glottal pulse spectral tilt LSFs ($g = 10$). For the sustained vowel database, we used the same input parameterization as for the continuous speech databases, but GIF was computed with a 100-ms frame length and a 100-ms frame skip.

From each complete training data set, 100,000 pulses were randomly selected for the DNN training. The used DNN architecture was a three hidden-layer feed-forward multilayer perceptron with hidden layer sizes of 250, 150, and 250 with sigmoid activations. The output layer size is 500 samples with a linear activation function. Each DNN was trained for 10,000 epochs.

4.3. Objective measures

The glottal pulse generation DNNs described in Section 4.2 were objectively evaluated using the average mean squared error (MSE) between the generated glottal sources and their corresponding references computed by QCP. In this evaluation, we used glottal pulses from test sets taken from the "Nick" and "Nancy" datasets. This arrangement divides DNNs to two categories: The "Nick", "Nancy", and "Nick+Nancy" DNNs contain training data from the target voice, which in principle should be reflected positively in the objective evaluation. For the rest of the DNN systems, the target voice is absent from the training data.

The average MSEs with their standard deviations for the performed tests are presented in Figure 3. The results suggest that in terms of MSE, voice-specific training data greatly increases the generation performance, as expected. Among the systems in which the target voice was absent from the training dataset, the sustained vowel database performs best, particularly for the male voice.

4.4. Subjective listening test

The glottal pulse generation DNNs were compared subjectively in analysis-synthesis type of a speech synthesis experiment with the GlottDNN vocoder. The subjective speech synthesis quality was studied with a MUSHRA (multiple stimuli with hidden reference and anchor [28]) test. The anchor samples were generated with a straightforward impulse-train excitation. As in the objective experiment, the "Nick" and "Nancy" voices were tested separately. To limit the test duration, the VCC-male and VCC-female DNNs were omitted. The listening test was implemented using the BeaqleJS framework [29], with 10 samples from both speakers. 14 expert listeners participated in the test.

The results of the subjective MUSHRA test with their 95% confidence intervals are presented in Figure 4. For the "Nancy" voice, the listening test results are surprising: The voice-specific glottal pulse generation DNN showed the worst performance out of the compared systems, with the other systems being mostly on par with each other. A possible explanation for this result and for its discrepancy from the objective evaluation shown in Figure 3 is that glottal pulses inverse filtered from the "Nancy" voice might have been distorted due to the challenging acoustical features (such as high f_0, rapid tempo etc.) of this voice (see Section 4.1.1). Even though the objective evaluation indicates that the DNN trained with the "Nancy" voice gives the lowest MSE error in generating glottal pulses from speech of "Nancy", these generated pulses do not give the best excitation waveform for the synthesis.

For the "Nick" voice, the voice-specific DNN is the best performing system, and the VC DNN shows clearly the worst performance. Among the DNNs in which the target speaker was absent from the training data, the "Sustained" DNN was again the best system.

5. Conclusions

In the present study, the DNN-based generation of glottal pulses was addressed by studying how data variety in the network training affects the generation both in terms of objective measures as well as in terms of subjective, analysis-synthesis speech quality. By comparing seven different training platforms, the study suggests that voice-specific training can greatly improve the performance of the glottal pulse generation if high-quality GIF estimates are available. Furthermore, if GIF quality is sacrificed in voice-specific training (e.g., due to having a voice with challenging acoustical features for GIF), the training data should be acquired from such voices that produce the best possible GIF estimates.

6. Acknowledgements

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

7. References

[1] P. Alku, "Glottal inverse filtering analysis of human voice production – A review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana*, vol. 36, no. 5, pp. 623–650, 2011.

[2] 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)*, 2014.

[3] J. Juvela, X. Wang, S. Takaki, M. Airaksinen, J. Yamagishi, and P. Alku, "Using text and acoustic features in predicting glottal excitation waveforms for parametric speech synthesis with recurrent neural networks," in *Proc. Interspeech*, 2014.

[4] 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.

[5] M. Airaksinen, B. Bollepalli, L. Juvela, Z. Wu, S. King, and P. Alku, "GlottDNN — A full-band glottal vocoder for statistical parametric speech synthesis," in *Proc. Interspeech*, 2016.

[6] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[7] M. Rothenberg, "A new inverse–filtering technique for deriving the glottal air flow waveform during voicing," *Journal of the Acoustical Society of America*, vol. 53, no. 6, pp. 1632–1645, 1973.

[8] M. Airaksinen, T. Raitio, and P. Alku, "Noise robust estimation of the voice source using a deep neural network," in *Proc. ICASSP*, 2015.

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

[10] 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.

[11] 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.

[12] 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.

[13] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Comparing glottal-flow-excited statistical parametric speech synthesis methods," in *Proc. ICASSP*, 2013.

[14] J. Flanagan, *Speech Analysis Synthesis and Perception*. Springer-Verlag Berlin Heidelberg, 1978.

[15] 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, pp. 1969–1973.

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

[17] 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.

[18] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. 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, 2012.

[19] C. Borß and R. Martin, "On the construction of window functions with constant-overlap-add constraint for arbitrary window shifts," in *Proc. ICASSP*, 2012, pp. 337–340.

[20] M. Bosi and R. Goldberg, *Introduction to Digital Audio Coding and Standards*. Springer US, 2003.

[21] J. J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, 2nd ed. Wiley-IEEE Press, 1999.

[22] M. Airaksinen, L. Juvela, T. Bäckström, and P. Alku, "Automatic glottal inverse filtering with non-negative matrix factorization," in *Proc. Interspeech*, 2016.

[23] D. Wong, J. Markel, and A. Gray Jr., "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 4, pp. 350 – 355, 1979.

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

[25] S. King and V. Karaiskos, "The blizzard challenge 2011," in *Blizzard Challenge 2011 Workshop*, 2011.

[26] 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

[27] T. Toda, L.-H. Chen, D. Satio, F. Villavencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016," in *Proc. Interspeech*, 2016.

[28] I. T. Union, "Itu-r bs.1534 (method for the subjective assessment of intermediate quality levels of coding systems)," 2015.

[29] S. Kraft and U. Zlzer, "BeaqleJS: HTML5 and JavaScript based-Framework for the Subjective Evaluation of Audio Quality," in *Linux Audio Conference*, 2014.