The NII speech synthesis entry for Blizzard Challenge 2016

Lauri Juvela[1], *Xin Wang*[2,3], *Shinji Takaki*[2], *SangJin Kim*[4], *Manu Airaksinen*[1], *Junichi Yamagishi*[2,3,5]

[1]Aalto University, Department of Signal Processing and Acoustics, Finland
[2]National Institute of Informatics, Japan
[3]Sokendai University, Japan
[4]Naver Labs, Naver Corporation, Korea
[5]University of Edinburgh, The Centre for Speech Technology Research, United Kingdom

lauri.juvela@aalto.fi, {wangxin,takaki,jyamagis}@nii.ac.jp

Abstract

This paper decribes the NII speech synthesis entry for Blizzard Challenge 2016, where the task was to build a voice from audiobook data. The synthesis system is built using the NII parametric speech synthesis framework that utilizes Long Short Term Memory (LSTM) Recurrent Neural Network (RNN) for acoustic modeling. For this entry, we first built a voice using a large data set, and then used the audiobook data to adapt the acoustic model to the target speaker. Additionally, the recent full-band glottal vocoder GlottDNN was used in the system with a DNN-based excitation model for generating glottal waveforms. The vocoder estimates the vocal tract in a band-wise manner using Quasi Closed Phase (QCP) inversefiltering at the low-band. At synthesis stage, the excitation model is used to generate voiced excitation from acoustic features, after which a vocal tract filter is applied to generate synthetic speech.

The Blizzard Challenge listening test results show that the proposed system achieves comparable quality with the benchmark parametric synthesis systems.

Index Terms: Blizzard Challenge, parametric speech synthesis, speaker adaptation, glottal vocoding, LSTM

1. Introduction

The TTS system for this entry is based on the NII statistical parametric speech synthesis framework, where the latest version of glottal vocoders [1] developed in Aalto University, the full-band glottal vocoder GlottDNN[2], is used instead of more conventional vocoding techniques. Acoustic modelling in our synthesis framework is based on Long Short-Term Memory (LSTM) Recurrent Neural networks (RNN), while HTS [3] is used for duration modeling. Additionally, the system uses a feedforward DNN-based glottal excitation model.

This year's task in Blizzard Challenge was to build a voice based on audiobook data read by a British English female speaker. While the data set is fairly large, the acoustic model typically needs even more data to benefit from the RNN architecture. For this reason we chose an adaptation approach, where the acoustic model is first trained on a large data set and then tuned with the target speaker data. Another issue was posed by the parametrization of the data, due to some reverberation and background noise being present in the recordings. After initial experiments with STRAIGHT [4] and WORLD [5] vocoders, we decided to use the current version of the GlottDNN vocoder.

Previously, female voices have been problematic for the glottal vocoding [6, 7], in contrast to the good results with male voices reported in [1, 2] in comparison with the STRAIGHT vocoder. However, recent improvements with a high-pitched voice in [8] encouraged us to use the new full-band glottal vocoder version to this voice building task. Since the vocoding method in this work is fairly novel, and no audiobook specific techniques were developed yet for the synthesis system, this paper focuses on giving detailed descriptions on the used vocoding and acoustic modeling techniques.

This paper is structured as follows: section 2 describes the data sets and pre-processing steps used for building the voice, while section 3 details the speech parametrisation and synthesis techniques, along with the acoustic and excitation models used. The results from the Blizzard Challenge listening tests will be discussed in section 4, with concluding remarks in section 5.

2. Data

2.1. Overview of the speech corpora

The data corpus released for the Blizzard Challenge this year consists of English audiobooks, all read by the same female speaker with a British accent. We utlize all the released data for system construction, including the pilot data released last year. In total, the uilized corpus contains 5729 utterances with the total duration of 300 minutes.

Because this audiobook corpus may not be sufficient to train the acoustic model based on deep neural network, we also utilize the *Nancy* corpus from Blizzard Challenge 2011 [9] to pre-train the neural network. This corpus contains 12092 utterances with a total duration of 963 minutes. Although this speaker has an American accent, the data set benefits from being specifically designed for speech synthesis and from being of high recording quality.

2.2. Speech data pre-processing

While the quality of recordings of the Nancy corpus is well controlled, the quality of the audiobook may not be ideal for parametric speech synthesis. Thus, pre-processing is conducted on the audiobook data as follows:

1. De-reverberation: the deverberator of Postfish [10] is utilized to take the unwanted room echo out of the speech recordings;
2. Noise reduction: the noise reduction function of Audacity [11] is used to attenuate the constant background noise in recordings. This function is in essence a multi-band digital noise gate, automatically shaped by the property noise extracted from a small segment of the recording;
3. Energy level normalization: the root mean square (RMS) energy level of all the recordings are normalized after voice activity detection and RMS level calculation.

2.3. Text data pre-processing

The text data was manually checked. First, because the text and audio segmentation was not always consistent, the text was manually checked so that the text matched the content of the speech. Second, non-speech content in the speech waveform was annotated in the text. The re-annotated text data were also shared with the Blizzard Challenge participants.

3. Speech synthesis system

An overview of the synthesis system back-end is shown in Figure 1. This section presents the procedure for feature extraction, acoustic and excitation model training, and speech waveform synthesis. The acoustic model uses LSTM RNN, while the glottal excitation model is a feed-forward DNN.

3.1. Feature extraction

3.1.1. Acoustic features

For other acoustic feature extraction and synthesis-time waveform generation we use the newly introduced full-band glottal vocoder, GlottDNN [2]. The vocoder extends the Quasi Closed Phase (QCP) [12] inverse-filtering analysis, and the DNN-based glottal excitation prediction presented in [8] to the 48 kHz sampling rate. Figure 2 shows the analysis program flow implemented in the GlottDNN vocoder, as explained in detail in [2]. The main vocoder property, regarding full-band spectral analysis, is the splitting of the speech signal into two frequency bands with Quadrature Mirror Filtering (QMF) [13]. With QMF, the signal is split into two frequency bands with mirrored frequency response filters and downsampled on both bands separately, resulting in half rate signals representing high and low frequency bands. This allows using the QCP analysis in the low-band, where the periodicity caused by the glottal excitation is more prominent, while using conventional linear prediction in the more aperiodic high-band. As a result, more parameters can be allocated to the perceptually more important lower frequencies. For modelling, the Line Spectral Frequency (LSF) representation is used for both vocal tract features.

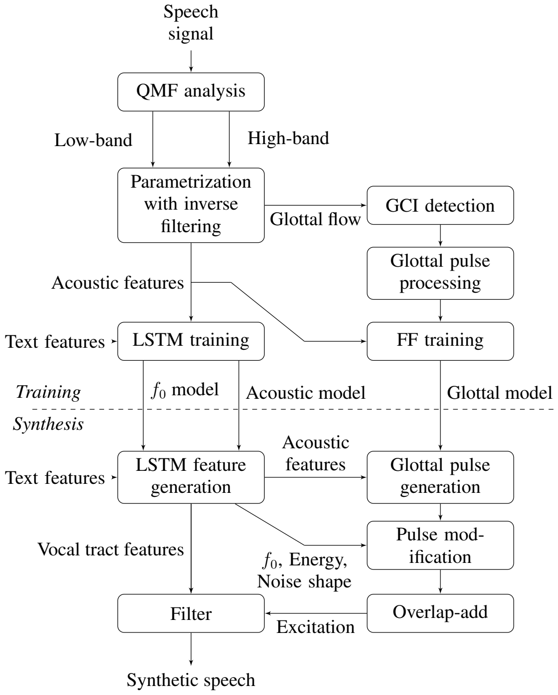

Figure 1: Synthesis system block diagram. At parametriation, the signal is split to high and low frequency bands, allowing different linear-predictive techniques for the bands. Glottal flow obtained from inverse-filtering is processed pitch-synchronously using the glottal closure instants (GCI), and a feed-forward (FF) DNN is trained to predict the glottal waveforms. The model predicting acoustic features from text is based on LSTM RNN.

Another novelty in the vocoder is in the modeling of the aperiodic component of the excitation signal. First, a glottal source estimate is obtained by inverse-filtering the speech signal with the combined vocal tract filter formed from the band-wise filters. Second, the glottal source is median filtered to obtain a noise-like residual that closely resembles the prediction residual of the DNN-based excitation model [2], and finally, the spectral shape of this noise signal is parametrized with line spectral frequency (LSF). The acoustic features and their dimensions are summarized in Table 1.

Due to the high expressiveness of the audiobook data, voting from several fundamental frequency (f_0) estimators was used for increased robustness. The extracted f_0 trajectory is based on the merged results of five f_0 extractors, comprising the glottal autocorrelation method [1], SWIPE [14], RAPT [15], SAC [16], and TEMPO [17]. Given the f_0 candidates of each frame, the median is selected as the f_0 observation. For the DNN acoustic model, a binary voiced/unvoiced decision (VUV) is separated from the f_0, and the f_0 trajectory is linearly interpolated to have a continuous value also at unvoiced regions. The median f_0 trajectory was also used in the vocoding.

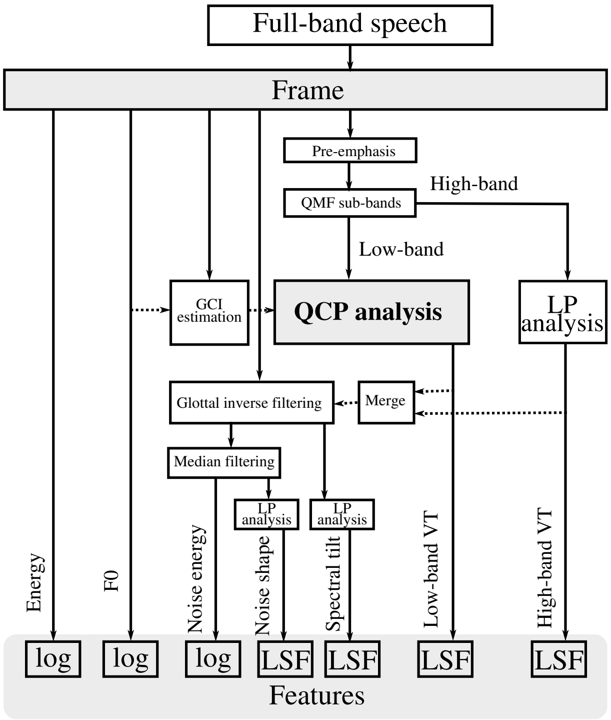

Figure 2: Vocoder analysis module block diagram [2]. Vocal tract (VT) filter analysis is performed in two frequency bands, where QCP is used on the low-band and conventional linear prediction (LP) is used on the high-band.

3.1.2. Text features

The system specifically uses two kinds of text features: first, the full-context phonetic labels for the HMM-based duration model, and second, the frame-rate input to the neural network acoustic model. The utilized text features are similar to those in the standard HTS system [3]. Because the neural network acoustic models are pre-trained on the Nancy data [9], the General American (GAM) accent of Combilex lexicon [18] was chosen as the phoneme set. For both the training and test data, the letter-to-sound conversion, part-of-speech tagging, syllable accent inferring, and Tone and Break Index (ToBI) intonational boundary tone prediction are all conducted by Flite [19]. The text features as input to the neural network also include the position of current frame in the phoneme and utterance. In this entry, passage or paragraph feature is not taken into consideration.

Table 1: Acoustic features and their dimensions (including their Δ and $\Delta\Delta$ values) used in the system. The first five acoustic features are utilized as the input to predict the glottal waveform.

Feature	dim.	Δ dim.
Fundamental frequency (log f_0)	1	3
Energy (log)	1	3
Low-band vocal tract (LSF)	42	126
High-band vocal tract (LSF)	18	54
Glottal source spectral tilt (LSF)	10	30
Voice-unvoiced decision (VUV)	1	1
Noise shape (LSF)	24	72
Noise energy (log)	1	3

3.2. Acoustic model training

The overview of the acoustic and excitation models in the synthesis system is shown in Figure 1. Left side of the figure depicts the model used to generate acoustic features from text-derived input features, and the right side shows the glottal excitation model used to generate glottal waveforms from acoustic features.

Differing from the glottal vocoding framework in [2], where one neural network is utilized for predicting the glottal waveform and another network for predicting all the acoustic features, the implemented framework in our system utilized an additional network to model the f_0 trajectory separately. Thus, there are in total three neural networks. This is motivated by our recent finding that a neural network may devote most of its network capacity to model the spectral features while assigning less priority to the perceptually important f_0 trajectory [20]. Note that instead of directly using the $\log f_0$, the f_0 trajectory is converted to mel-scale with the relation $m = 1127 * \log(1 + f_0/700)$ where f_0 is the fundamental frequency in Hz [21].

The neural networks for predicting f_0 and other acoustic features are implemented based on the RNN with bi-directional LSTM units. For the f_0 trajectory prediction, the neural network is constructed with two feedforward layers near the input side, followed by two

LSTM layers. The layer size of the feedforward layers is set to 1024, while the size of the LSTM layers is 512.

The training stage of the acoustic model consists of two steps. First, the network is randomly initialized and trained given the data from the Nancy corpus. 500 sentences from this corpus are utilized as the validation set and the rest of the data are used for training. Stochastic gradient descent with early stopping is adopted. Given the network trained on the Nancy data, the second step is to fine-tune the network using the audiobook data of the current task. The training process for the second step is similar to the first step, except the size of the validation set is 200.

The duration model at the phoneme level, which is not shown in Figure 1, uses a fairly standard HMM-based parametric HTS framework. The decision-tree-based model clustering process results in 2087 clustered models out of the 162781 full-context models.

3.2.1. Excitation model

The synthesis system utilizes a DNN-based excitation model that predicts glottal excitation waveforms from the features generated by the acoustic model. This concept was first introduced in [22], while this paper follows the waveform processing method presented in [8]. For training the model, glottal pulses are extracted from the signal estimated by inverse-filtering, as illustrated in Figure 3. First, glottal closure instants (GCI), defined as the periodic minima in the glottal flow derivative waveform, are detected. Using the GCI, two pitch-period glottal pulses are extracted, cosine windowed, and zero padded to a desired fixed length. In this case, pulse length of 1600 was chosen, corresponding to a minimum f_0 of 60 Hz.

The network for modelling the glottal waveform is implemented with a fully connected feedforward neural network. The input features include the first five kinds of acoustic features listed in Table 1, i.e. the noise features and the binary VUV decision are excluded. The output is the feature vector corresponding to the 1600 sampling points of the glottal waveform. This network consists of 4 hidden layers (with sizes 250, 100, 250 and 1400), and each layer utilizes the sigmoid activation function. The excitation model was trained using data only from the target speaker.

3.3. Speech synthesis

At the synthesis front end, the input text is first split into sentence-length segments, as the current text-to-phonetic-labels system only handles context up to the sentence level. The paragraph level text segments required for testing are simply concatenated from the individual sentences after synthesis. For the sentence-level text inputs, Flite is used to create phonetic labels from the input. HTS-based duration model trained on the target speaker is used with the Combilex lexicon to create the frame-rate text features for the neural network inputs.

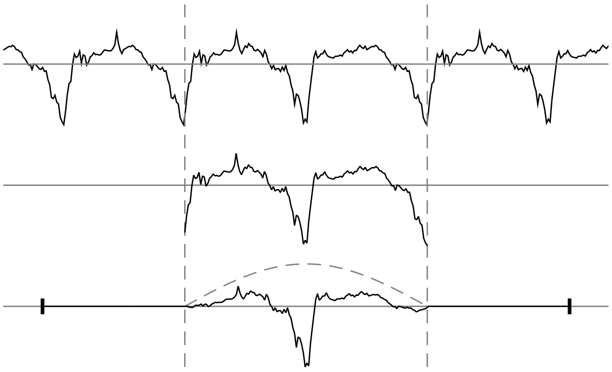

Figure 3: Glottal exitation pulses are formatted for DNN by taking a two pitch-period segment delimited by GCI, cosine windowing the pulse, and zero-padding to a fixed length.

At the synthesis back-end, the text feature vectors are used to generate the dynamic acoustic features listed in Table 1. The Maximum Likelihood Parameter Generation (MLPG) algorithm is utilized to create smooth feature trajectories, and the resulting features are used for both the input of the excitation model and final waveform generation with the vocoder. An overview of the synthesis system back-end is shown in Figure 1, while the vocoder synthesis procedure is detailed in Figure 4.

The waveform synthesis process is done similarly to [2]: first, the voicing decision is determined from the f_0, and in the voiced case the acoustic features are fed into the excitation DNN to create glottal excitation pulses. These pulses are first truncated to match the generated f_0 and cosine windowed, summing up to a Hann window, which is required for the overlap-add procedure. The pulses are then modified for aperiodicity by adding a noise component based on the noise shape LSFs, after which spectral matching is applied to compensate any difference between predicted spectral tilt and generated pulse spectrum. As a final modification, the pulses are scaled by energy. The modified pulses are then assembled into a voiced excitation signal with the pitch-synchronous overlap-add method [23], using synthesis pitch marks determined by the f_0. Unvoiced excitation is simply created by scaling white noise to the desired energy level. Finally, the vocal tract filter is merged from the generated high-band and low-band LSFs and used to filter the excitation, resulting in synthetic speech.

4. Results and analysis

The synthesis system was evaluated as a part of the Blizzard Challenge listening tests, where the participating entries were evaluated by speech experts and paid listeners in controlled listening conditions, and online volunteers

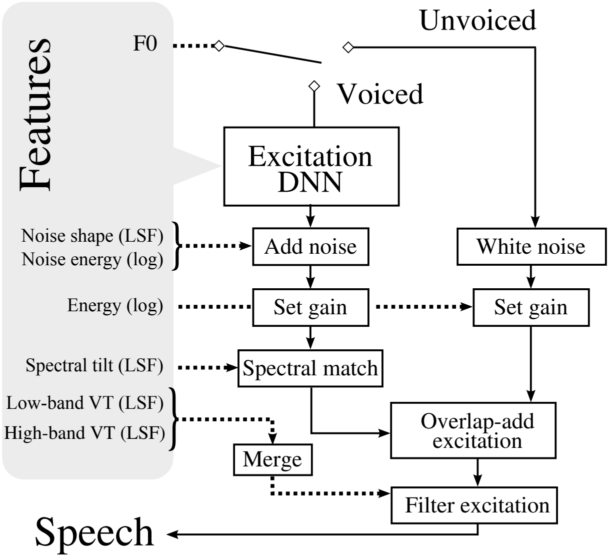

Figure 4: Vocoder synthesis module block diagram [2].

in varying conditions. Here we focus on pooled results from all listeners to get a general impression of the results.

Figure 5 shows the naturalness ratings presented as box plots, where the central solid bar marks the median, the shaded box represents the quartiles, and the extended dashed lines show 1.5 times the quartile range. The most relevant comparisons can be made with the other known parametric synthesis systems, namely system C, which is the HTS benchmark system, and system D, which a DNN benchmark build with the new toolkit by CSTR (University of Edinburgh). The results show that our proposed system K outperforms the HTS benchmark and ranks similarly with the DNN benchmark. Wilcoxon signed rank tests further indicate that the difference between the proposed system and HTS benchmark is statistically significant, whereas the difference to the DNN benchmark is not significant.

Speaker similarity scores are presented in Figure 6 with similar box plots. The results show that the proposed system has comparable level of speaker similarity to the HTS benchmark, while having lower similarity than the DNN benchmark. This is supported by the significance tests, which indicate no significant difference between the proposed system and HTS benchmark. Two possible reasons may have lead to the relatively low similarity score. First, the acoustic model was pre-trained using the Nancy data; second, the GAM American accent phoneme set was used for the target speaker, whose accent is different.

5. Conclusion

Although parametric synthesis is generally not yet as good as unit-selection synthesis, a positive finding from

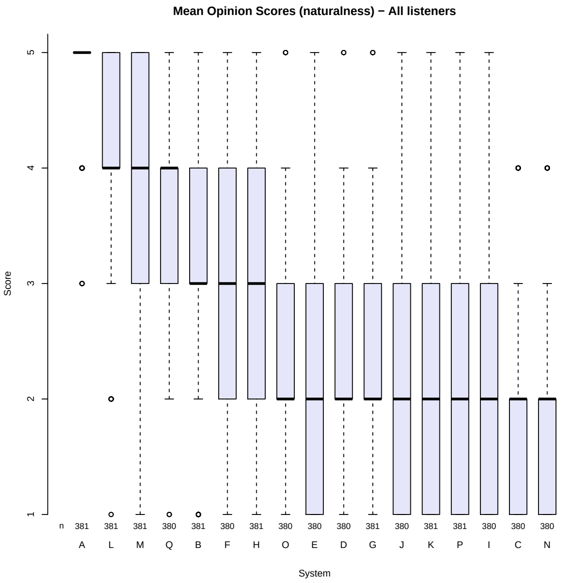

Figure 5: Naturalness ratings. System K is the proposed system, C is the HTS benchmark, and D is the DNN benchmark.

the glottal vocodings perspective in the present study was that we achieved a similar performance to known benchmark parametric systems. It is worth emphasizing that this happened even though the synthesis was based on a female voice, which is known to be challenging speech data for glottal inverse-filtering analysis [6, 7]. Building this system also furthered the development of the new GlottDNN vocoder and DNN-based voice adaptation.

We feel that the audiobook data set was challenging for parametric synthesis, partially due to the expressiveness inherent to audiobooks, but also because of the signal level non-idealities affecting vocoding. In the future, more attention should be given to data preprocessing, namely experimenting more with state-of-the-art de-reverberation and noise suppression methods, and applying a more strict speech/non-speech classification, as the audiobook data also contained non-speech signals such as ambient effects.

6. Acknowledgements

The reseach in this paper was supported by the Naver Corp., the Academy of Finland (proj. no. 256961 and 284671), the European Union TEAM-MUNDUS scholarship (TEAM1400081), and partially supported by EPSRC through Programme Grants EP/I031022/1 (NST) and EP/J002526/1 (CAF), by the Core Research for Evolutional Science and Technology (CREST) from the Japan Science and Technology Agency (JST) (uDialogue project), and by MEXT KAKENHI Grant Numbers (26280066, 26540092, 15H01686, 15K12071,

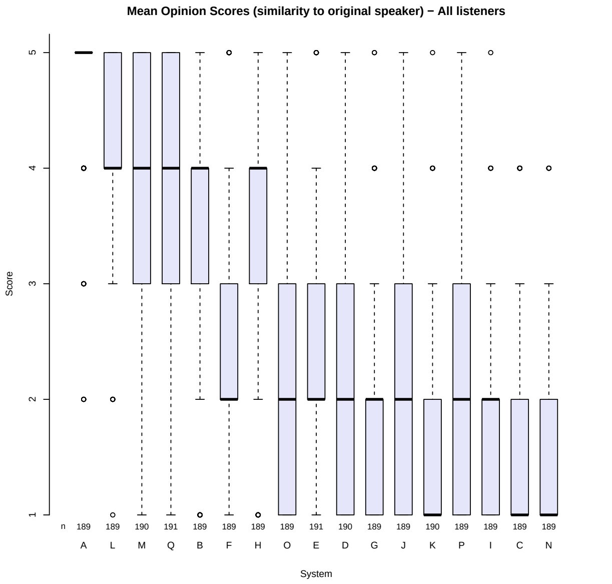

Figure 6: Speaker similarity ratings. System K is the proposed system, C is the HTS benchmark, and D is the DNN benchmark.

16H06302).

7. References

[1] 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.

[2] 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 *Interspeech*, Sept. 2016, pp. –.

[3] K. Tokuda, H. Zen, and A. W. Black, “An HMM-based speech synthesis system applied to english,” in *IEEE Speech Synthesis Workshop*, 2002, pp. 227–230.

[4] H. Kawahara, I. Masuda-Katsuse, and A. 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, pp. 187–207, 1999.

[5] M. Morise, T. Nishiura, and H. Kawahara, “Proposal of WORLD, a high-quality voice analysis, manipulation and synthesis system and its evaluation,” *ASJ technical report (in Japanese)*, vol. 41, no. 7, pp. 555–560, oct 2011.

[6] 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*, Turin, Italy, September 2011.

[7] ——, “The GlottHMM entry for Blizzard Challenge 2012: Hybrid approach,” in *Blizzard Challenge 2012 Workshop*, Portland, Oregon, September 2011.

[8] 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.

[9] S. King and V. Karaiskos, “The Blizzard Challenge 2011,” in *Blizzard Challenge 2011 Workshop*, Turin, Italy, September 2011.

[10] M. Montgomery, “Postfish by Xiph.org,” 2005. [Online]. Available: https://svn.xiph.org/trunk/postfish/README

[11] D. Mazzoni and R. Dannenberg, “Audacity,” 2000–2015. [Online]. Available: http://www.audacityteam.org/download/

[12] 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.

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

[14] A. Camacho, “SWIPE: A sawtooth waveform inspired pitch estimator for speech and music,” Ph.D. dissertation, University of Florida, 2007.

[15] D. Talkin, “A robust algorithm for pitch tracking (RAPT),” *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[16] E. Gómez and J. Bonada, “Towards computer-assisted flamenco transcription: An experimental comparison of automatic transcription algorithms as applied to a cappella singing,” *Computer Music Journal*, vol. 37, no. 2, pp. 73–90, 2013.

[17] H. Kawahara, A. de Cheveigné, and R. D. Patterson, “An instantaneous-frequency-based pitch extraction method for high-quality speech transformation: revised TEMPO in the STRAIGHT-suite.” in *ICSLP*, 1998.

[18] K. Richmond, R. Clark, and S. Fitt, “On generating combilex pronunciations via morphological analysis,” in *Interspeech*, 2010.

[19] HTS Working Group, “The English TTS System ”Flite+hts_engine”,” 2014. [Online]. Available: http://hts-engine.sourceforge.net/

[20] X. Wang, S. Takaki, and J. Yamagishi, “Investigating very deep highway networks for parametric speech synthesis,” in *SSW-9*, 2016.

[21] D. O’Shaughnessy, *Speech communications: human and machine*. Institute of Electrical and Electronics Engineers, 2000.

[22] 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.

[23] E. Moulines and F. Charpentier, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,” *Speech communication*, vol. 9, no. 5-6, pp. 453–467, 1990.