NOISE ROBUST ESTIMATION OF THE VOICE SOURCE USING A DEEP NEURAL NETWORK

Manu Airaksinen, Tuomo Raitio, Paavo Alku

Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland

ABSTRACT

In the analysis of speech production, information about the voice source can be obtained non-invasively with glottal inverse filtering (GIF) methods. Current state-of-the-art GIF methods are capable of producing high-quality estimates in suitable conditions (e.g. low noise and reverberation), but their performance deteriorates in non-ideal conditions because they require noise-sensitive parameter estimation. This study proposes a method for noise robust estimation of the voice source by creating a mapping using a deep neural network (DNN) between robust low-level speech features and the desired reference, a time-domain glottal flow computed by a GIF method. The method was evaluated with two GIF methods, of which one (quasi closed phase analysis, QCP) requires additional parameter estimation and the other (iterative adaptive inverse filtering, IAIF) does not. The results show that the proposed method outperforms the QCP method with SNRs less than 50–20 dB, but the simple IAIF method only with very low SNRs.

Index Terms— Voice source estimation, glottal inverse filtering, deep neural network, noise robustness

1. INTRODUCTION

In the production of voiced speech, the quasiperiodic fluctuation of the vocal folds generates an input signal to the vocal tract. In acoustical terms, this excitation signal is referred to as the glottal volume velocity waveform or the glottal flow. More generally, the waveform is called the voice source or the glottal source. This signal carries information about the type of phonation and pitch that can be associated with various vocal cues corresponding to the speaker, e.g. emotional state, individual speech characteristics, and possible voice pathologies. Unfortunately the real glottal flow is elusive to direct acoustical measurement due to the hidden location of the vocal folds inside the larynx. By using glottal inverse filtering (GIF), however, the glottal source can be estimated non-invasively from the speech pressure waveform recorded by a microphone outside the lips. GIF methods operate by applying such anti-resonances to a segment of recorded speech that the effects of the vocal tract formants become canceled, hence yielding an estimate of the glottal flow.

Examples of known GIF methods are closed phase covariance analysis (CP) [1], iterative adaptive inverse filtering (IAIF) [2], and complex cepstral decomposition (CCD) [3] (for more details, see recent reviews [4–6]). All the above methods, however, have been designed to work in ideal conditions in which the speech signal to be analyzed is typically a long sustained vowel produced in an environment with minimal noise and reverberation. Experiments conducted in ideal laboratory conditions, however, cannot be generalized to more realistic scenarios in which the glottal flow is to be estimated, for example, from continuous noisy speech. In these realistic scenarios, the accuracy of most GIF methods unfortunately deteriorates, and this holds true especially for the (theoretically) more powerful state-of-the-art methods such as CP, CCD, and quasi closed phase analysis (QCP) [7], which require precise estimates for glottal closure instants (GCIs). More straightforward methods, such as IAIF, do not require additional parameter estimation, which makes their performance more robust to noise. This in turn makes simple GIF methods applicable in modern data driven applications, such as statistical parametric speech synthesis [8–10], which call for estimation of the glottal flow from continuous speech signals that might have been recorded in non-ideal conditions. Enhancing the robustness of the state-of-the-art GIF methods could thus in principle result in a better performance in practical applications, particularly for techniques, such as statistical speech synthesis, in which glottal source estimates are used with model training from long speech recordings.

This study aims to enhance the robustness of the state-of-the-art GIF methods in low signal-to-noise ratio (SNR) conditions by creating a multi-speaker mapping between robust low-level speech parameters and the output of a reference GIF method. The mapping is done by utilizing a deep neural network (DNN), which is a powerful tool for finding nonlinear interactions between input and output features, even if the data is highly correlated [11]. The motivation for the study is to alleviate the effects caused by unreliable parameter estimation in the state-of-the-art GIF methods in low SNR conditions hence enabling more accurate estimation of the glottal flow in non-ideal conditions. The DNN-based approach to glottal source modeling and estimation is relatively little explored, which is the reason why this study should most of all be treated as a proof-of-concept. Background information on the research topic and the selected GIF methods are presented in Section 2, and the proposed method is explained in detail in Section 3. The used speech database, experimental setup, and results are detailed in Section 4. Finally, summary of the findings and discussion are provided in Section 5.

2. BACKGROUND

In automatic speech recognition (ASR), DNNs have provided significant improvements in recognition accuracy compared to previous state-of-the-art methods [11]. Also in text-to-speech (TTS) synthesis, recent studies utilizing deep learning architectures have provided promising results (e.g. [12, 13]). These improvements are enabled by the ability of a deep learning architecture to model complex dependencies between input and output features and utilize correlated high-dimensional data [11].

In ASR and TTS, a mapping is created between acoustic and linguistic features. In recent work on voice source modeling in statistical parametric speech synthesis [14, 15], a similar mapping using DNN is created between acoustic speech features and glottal flow

This work has been supported by the EC-FP7 (2007–2013) n° 287678 (Simple^{4}All) and the Academy of Finland (256961).

time-domain waveform. The results in [14,15] indicate that the glottal flow waveform can be successfully predicted from higher-level speech features, and the DNN-based system was rated equal to a high-quality baseline system in subjective listening tests. The input features in [14, 15] included, e.g., spectrum, gain, and fundamental frequency of speech, which all contain information on the voice source, and can be rather easily and robustly estimated from a speech signal. The DNNs trained in those two studies were speaker dependent, but a speaker-independent voice source DNN was trained and successfully used for synthesis in [16]. Utilizing the same principle, it is possible to predict the glottal flow signal from any speech frame using a speaker-independently trained DNN, which may be useful also in domains outside TTS.

In the study by Kane et al. [17], artificial neural network was successfully used for estimating the open quotient for different voice qualities using spectral features of speech. In the current study, the aim is to estimate the entire time-domain glottal flow waveform. Although the glottal flow estimate predicted by the DNN might not be as accurate as the one computed by a state-of-the-art GIF method (e.g. [7]), there are benefits in the afore-mentioned approach. First, the glottal flow estimate can be predicted using simple and robust speech features, which enables glottal flow estimation even if the speech frame is corrupted by noise. Especially more complex glottal inverse filtering methods that are subject to vulnerable parameter estimation (e.g., extraction of GCIs) suffer from noise [18] and other distortions (such as phase distortion [19]). Secondly, the estimation of the entire glottal flow waveform instead of only few descriptive parameters may have more practical applications, as shown in speech synthesis [14–16].

In this study, the IAIF [2] and QCP [7] methods were selected for estimating the initial glottal flow. IAIF is a widely known GIF method that works by in turn obtaining more accurate estimates of the spectral shape of the glottal flow and the vocal tract transfer function. It does not require additional parameter estimation, so it is selected for the study on the grounds of being a noise-robust baseline method. The QCP method is shown to provide better glottal flow estimates than IAIF [7], but requires additional parameter estimation that is sensitive to noise.

3. PROPOSED METHOD

The flow chart of the proposed method is illustrated in Figure 1. First, a large multi-speaker speech database is required in order to enable the glottal flow estimation for any speaker with a reasonable accuracy. The amount of data is also beneficial for the DNN training, which usually performs better when the amount of data is increased. Next, an existing glottal inverse filtering method is applied to the speech database to estimate the glottal flow of voiced speech. The glottal flow signal is segmented to individual two-pitch-period glottal flow segments, which are resampled to a constant length, windowed using the Hann window, and normalized in energy. Speech features extracted from the database are then linked with the corresponding glottal flow segments, and a mapping is established by training a DNN. The speech features can include any parametric representation of speech that enable predicting the glottal flow waveform. The most obvious choices are, e.g., spectral information, fundamental frequency, and frame energy, which all contain information from the voice source.

The DNN consists of the input layer, two hidden layers, and an output layer. The size of the input and output layers are defined by the dimensions of the input feature vector and the size of the resampled glottal flow waveform, respectively. With 16 kHz sampling

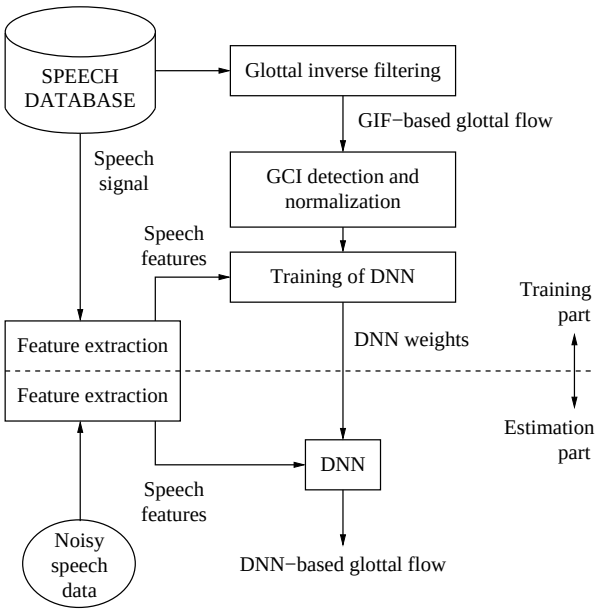

Fig. 1. Illustration of the proposed method.

rate, the length of the glottal flow segment is set to 400 samples, which results in 400 neurons in the output layer. The DNN architecture is based on the experiments conducted in [15,16], where 200 neurons in the hidden layers was found to perform best. Sigmoid activation functions are used in the hidden layers and linear activation functions in the output layer. The network weights are initialized by random Gaussian numbers with zero mean and standard deviation of 0.1. The network is then trained using back-propagation. The DNN code is based on [20] and modified for the purpose of the study.

After the DNN training is converged, the DNN network can be used to generate estimates of the glottal flow using new, possibly noisy speech data. The same feature extraction is applied to the new speech signal, and the extracted features are then fed to the trained DNN, which finally outputs the glottal flow estimate.

4. EXPERIMENTS

The experiments conducted in this study evaluate the SNR dependent performance of the proposed method using various input speech features. The performance, i.e., the accuracy of the glottal flow estimation, is compared to the initial glottal flow estimates obtained with the two selected inverse filtering methods, IAIF and QCP. All implementations of the proposed method were trained and tested using a large multi-speaker corpus of high-quality speech recordings.

4.1. Speech data

A multi-speaker speech database was constructed for the study. Ten high-quality male speech databases, all designed for speech synthesis purposes, were used as the speech data. Only male speech was selected for this preliminary study since the estimation of the voice source from female speech is generally more difficult than from male speech. Also training a mixed-gender DNN may result in less accurate results due to differences in male and female voice source characteristics. Altogether, the speech database consisted of 11 042 sentences, comprising about 17.5 hours of speech data. The languages and number of sentences of each speaker in the database are shown

Table 1. Details of the speech data.

Speaker	Gender	Language	Sentences	Length (min)
1	Male	Finnish	429	27
2	Male	Assamese	1466	127
3	Male	English	1138	80
4	Male	English	1131	51
5	Male	Gujarati	450	123
6	Male	Hindi	875	121
7	Male	Finnish	692	67
8	Male	English	2022	134
9	Male	Rajasthani	1369	133
10	Male	Telugu	1470	187
			11 042	1050 = 17.5 h

in Table 1. All speech was sampled at 16 kHz. Before feature extraction and glottal inverse filtering, the loudness of the speech files was normalized using the method in ITU-T P.56 [21]. The polarity of each database was checked and corrected in case it was inverted in order to guarantee correct modeling of the glottal flow waveform.

4.2. Experimental setup

The speech data was analyzed using 30 ms frames at 15 ms intervals, resulting in a total number of 1 917 832 voiced frames. The frames were divided into a training set consisting of 98% of the total number of frames, and a test set containing the remaining 2%. The frames were inverse filtered with the IAIF [2] and QCP [7] methods and input feature extraction was performed on the frames according to the desired system setups. Fundamental frequency (f_0) and frame energy (E) were included in all test systems (for MFCCs in the form of MFCC-0), along with a varying spectral feature representation. The selected spectral features contained varying parameter orders of line spectral frequencies (LSFs) [22], mel-frequency cepstral coefficients (MFCCs), and also GlottHMM vocoder [9, 10] vocal tract and voice source LSF estimates. The total number of different input vector combinations was 7, resulting in a total number of 14 systems for the whole experiment. A detailed list of the evaluated input and output parameter combinations is presented in Table 2.

In the experiment, white Gaussian noise was added to the test set frames according to the desired SNR. The SNR was varied from virtually clean speech (80 dB) to highly corrupted speech (0 dB) using 10 dB steps. Clean and corrupted inverse filtering estimates were then obtained with both GIF methods. The DNN input parameters were computed from the corrupted frames and fed into the DNN to obtain the estimates, which were then compared to the clean signal estimates of the respective methods.

4.3. Evaluation methods

The accuracy of glottal flow estimation in comparison to the clean reference was measured using four metrics. The normalized amplitude quotient (NAQ) [23] was used, which is a widely used and robust measure of voice quality. The magnitude difference between the first and the second harmonics, denoted as H1H2 [24], was used for measuring the performance in the spectral domain. The mean squared error (MSE) between the initial glottal flow estimate and the one predicted by the DNN was also measured, which is maybe the most important measure here, since the objective of the proposed method is to estimate the entire glottal flow signal. Finally, spectral distortion (SD) [25] was also measured, which is a widely used distortion measure that quantifies distortion along all frequencies.

Table 2. Evaluated input parameters and the GIF method used in the computation of the output glottal flow estimate.

System	Input	Output
IAIF-GlottHMM	f_0, E, 30×LSFvt, 10×LSFs	IAIF
IAIF-LSF18	f_0, E, 18×LSF	IAIF
IAIF-LSF30	f_0, E, 30×LSF	IAIF
IAIF-LSF46	f_0, E, 46×LSF	IAIF
IAIF-MFCC13	f_0, 13×MFCC	IAIF
IAIF-MFCC30	f_0, 30×MFCC	IAIF
IAIF-MFCC46	f_0, 46×MFCC	IAIF
QCP-GlottHMM	f_0, E, 30×LSFvt, 10×LSFs	QCP
QCP-LSF18	f_0, E, 18×LSF	QCP
QCP-LSF30	f_0, E, 30×LSF	QCP
QCP-LSF46	f_0, E, 46×LSF	QCP
QCP-MFCC13	f_0, 13×MFCC	QCP
QCP-MFCC30	f_0, 30×MFCC	QCP
QCP-MFCC46	f_0, 46×MFCC	QCP

Table 3. Average errors over all SNRs for the 14 systems in Table 2 and for the two reference GIF methods, IAIF and QCP. NAQ error is relative, H1H2 and SD errors are in dB, MSE is absolute. Smallest errors of the test systems are highlighted with bold font.

System	NAQ	H1H2	MSE	SD
IAIF-Ref	0.10	1.32	0.34	4.20
IAIF-GlottHMM	0.25	2.32	0.44	7.11
IAIF-LSF18	0.25	2.32	0.44	6.68
IAIF-LSF30	0.26	2.27	0.44	**6.53**
IAIF-LSF46	0.26	**2.26**	0.44	6.36
IAIF-MFCC13	0.24	2.35	0.43	6.58
IAIF-MFCC30	**0.20**	2.38	**0.40**	6.71
IAIF-MFCC46	0.22	2.35	**0.40**	6.87
QCP-Ref	0.18	1.64	0.50	5.63
QCP-GlottHMM	0.29	2.43	0.45	5.98
QCP-LSF18	0.23	2.41	0.43	6.18
QCP-LSF30	0.25	2.35	0.43	5.92
QCP-LSF46	**0.22**	2.37	0.43	6.34
QCP-MFCC13	0.24	**2.29**	0.43	**5.75**
QCP-MFCC30	0.24	2.37	0.43	5.94
QCP-MFCC46	0.23	2.47	**0.42**	6.01

4.4. Results

The obtained average results are shown in Table 3, and a detailed graph of the results is presented in Figure 2. All of the tested systems are able to reproduce the output of the reference method with very similar accuracy compared to each other. Figure 2 also illustrates that the error of the DNN output stays very stable with varying SNR, which suggests that the DNN-based systems are robust to noise. However, even though the errors for the IAIF and QCP based systems are similar, the proposed method is only truly advantageous for the QCP method for SNR cases below 50–20 dB, depending on the used error metric, as those are the levels where the errors of the proposed method become smaller than the errors for the reference method. For IAIF, most of the metrics give errors smaller than the baseline reference only for very low SNRs from around 0 to 10 dB. These results are in line with the initial speculations on the performance of the proposed method, i.e., the method requiring noise-sensitive parameter estimation can benefit from the DNN mapping, whereas the more simple method is noise-robust in its own

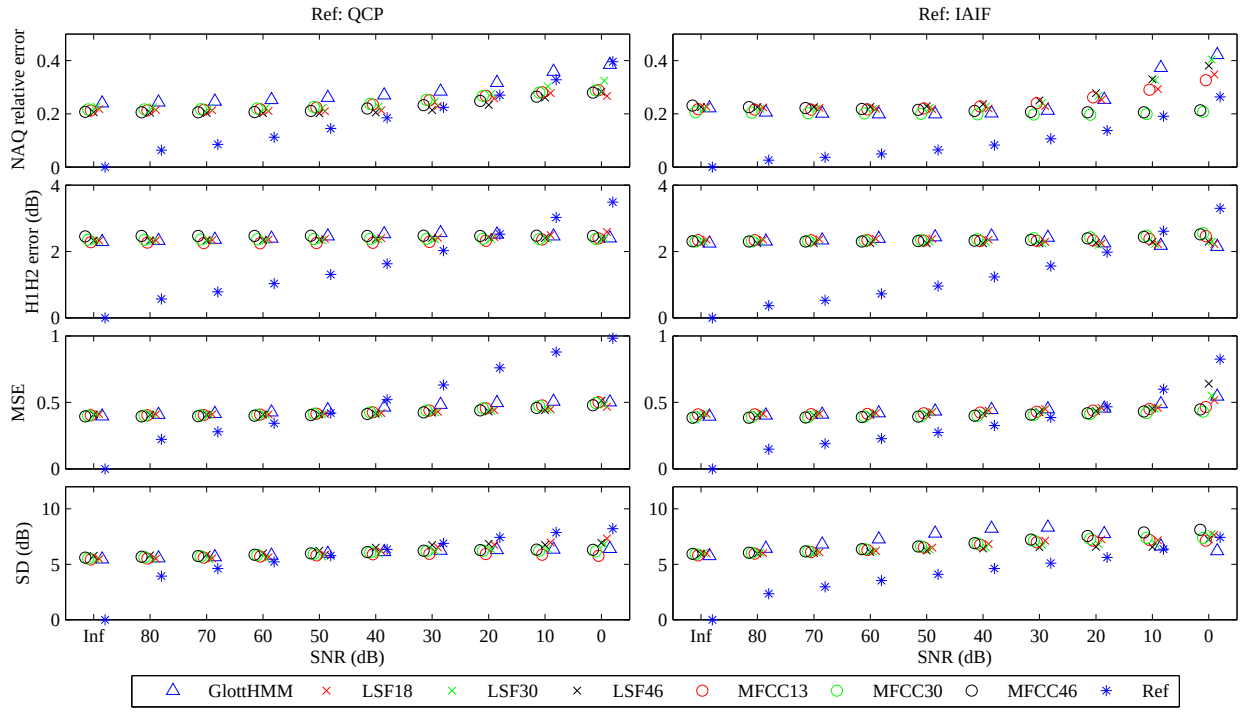

Fig. 2. Results of the evaluation for the NAQ, H1H2, MSE, and SD measures for the 14 different systems and the two reference GIF methods. The left graphs show the results for the QCP reference method while the right graph shows the results using the IAIF as the reference.

right.

An illustration on the ability of the proposed method to produce noise robust glottal flow estimates is presented in Figure 3, where the output of the QCP-LSF46 system is compared to the reference outputs of the QCP and IAIF methods with varying SNR. The DNN-based method outputs a slightly averaged glottal flow waveform, missing some of the finer details of the reference waveform with higher SNRs. However, in this example the shape of the reference QCP waveform starts to deteriorate starting from SNR of 40 dB, after which the deterioration is very severe. Meanwhile, the DNN output is very consistent until 10 dB SNR, after which the overall shape starts to slightly shift, but still preserving the shape of a glottal flow derivative waveform. The IAIF reference deteriorates less than the QCP reference, but also with very low SNRs, the glottal flow shape is severely distorted, while the DNN output maintains a consistent shape.

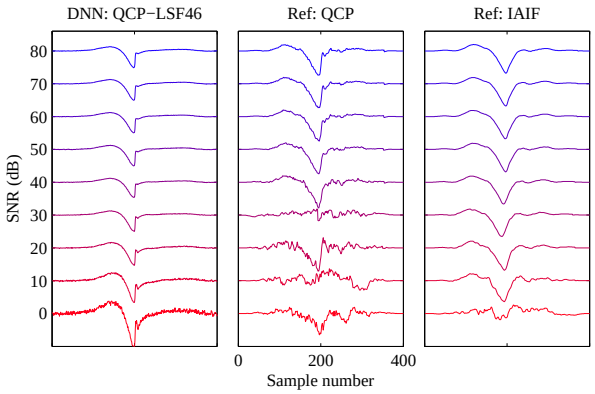

Fig. 3. Example glottal flow waveforms generated by the proposed DNN-based method (QCP-LSF46), and the QCP and IAIF methods using varying SNR. The DNN-based method clearly gives more consistent results with decreasing SNR.

5. DISCUSSION

The experiments show that the proposed DNN-based voice source estimation method yields noise robust estimates of the glottal flow signal. However, the commonly used voice quality metrics, such as NAQ and H1H2, show relatively high errors at all SNRs, even with a clean signal. This is due to the averaging effect of the DNN, which, for example, outputs pulses with slightly over-smooth waveform at the GCI, which in the natural glottal flow signal shows a rather abrupt discontinuity. Since NAQ is higly sensitive to the abruptness of the signal at the GCI, it is obvious that the errors of the DNN-based methods measured with NAQ are rather high. Despite the relatively high errors in these conventional voice quality metrics, the MSE instead shows that the glottal flow signal is very close to the clean reference estimate. With lower SNRs, the DNN mapping is very robust to noise, being able to yield reasonable glottal flow estimates when other methods give very distorted output (see Figure 3). Moreover, the studies in [14–16] show that despite the somewhat over-smooth characteristics of the glottal flow output from the DNN, the generated glottal flow signal is useful and feasible in speech synthesis. This suggests that estimating the entire glottal flow signal instead of predicting only a few descriptive parameters (such as in [17]) may be useful in other applications as well.

6. REFERENCES

[1] D. Wong, J. Markel, and A. Gray Jr., "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 27, no. 4, pp. 350–355, 1979.

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

[3] T. Drugman, B. Bozkurt, and T. Dutoit, "Complex cepstrum-based decomposition of speech for glottal source estimation," in *Proc. Interspeech*, 2009, pp. 116–119.

[4] J. Walker and P. Murphy, "Advanced methods for glottal wave extraction," in *Nonlinear Analyses and Algorithms for Speech Processing*, M. Faundez-Zanuy et al., Eds., pp. 139–149. Springer Berlin/Heidelberg, 2005.

[5] 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.

[6] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech & Language*, vol. 28, no. 5, pp. 1117–1138, 2014.

[7] 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.

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

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

[10] 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. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4564–4567.

[11] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Sig. Proc. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[12] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7962–7966.

[13] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Proc. Interspeech*, 2014, pp. 7962–7966.

[14] 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.

[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] A. Suni, T. Raitio, D. Gowda, R. Karhila, M. Gibson, and O. Watts, "The Simple4All entry to the Blizzard Challenge 2014," in *Blizzard Challenge 2014 Workshop*, 2014.

[17] J. Kane, S. Scherer, L.-P. Morency, and C. Gobl, "A comparative study of glottal open quotient estimation techniques," in *Proc. Interspeech*, 2013, pp. 1658–1662.

[18] T. Drugman, B. Bozkurt, and T. Dutoit, "A comparative study of glottal source estimation techniques," *Computer Speech & Language*, vol. 25, no. 1, pp. 20–34, 2012.

[19] J. N. Holmes, "Low-frequency phase distortion of speech recordings," *J. Acoust. Soc. Am.*, vol. 58, no. 3, pp. 747–749, 1975.

[20] G. Hinton, "Training a deep autoencoder or a classifier on MNIST digits," http://www.cs.toronto.edu/~hinton/MatlabForSciencePaper.html, last visited Oct. 2014.

[21] ITU, "Objective measurement of active speech level," International Telecommunication Union, Recommendation ITU-T P.56, 2011.

[22] F. K. Soong and B.-H. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1984, vol. 9, pp. 37–40.

[23] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parameterization of the glottal flow," *J. Acoust. Soc. Am.*, vol. 112, no. 2, pp. 701–710, 2002.

[24] I. Titze and J. Sundberg, "Vocal intensity in speakers and singers," *J. Acoust. Soc. Am.*, vol. 91, no. 5, pp. 2936–2946, 1992.

[25] F. Nordin and T. Eriksson, "A speech spectrum distortion measure with interframe memory," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2001, vol. 2, pp. 717–720.