Deep neural network based trainable voice source model for synthesis of speech with varying vocal effort

Tuomo Raitio[1], *Antti Suni*[2], *Lauri Juvela*[1], *Martti Vainio*[2], *Paavo Alku*[1]

[1]Department of Signal Processing and Acoustics, Aalto University, Finland
[2]Institute of Behavioural Sciences, University of Helsinki, Finland
firstname.lastname@aalto.fi, firstname.lastname@helsinki.fi

Abstract

This paper studies a deep neural network (DNN) based voice source modelling method in the synthesis of speech with varying vocal effort. The new trainable voice source model learns a mapping between the acoustic features and the time-domain pitch-synchronous glottal flow waveform using a DNN. The voice source model is trained with various speech material from breathy, normal, and Lombard speech. In synthesis, a normal voice is first adapted to a desired style, and using the flexible DNN-based voice source model, a style-specific excitation waveform is automatically generated based on the adapted acoustic features. The proposed voice source model is compared to a robust and high-quality excitation modelling method based on manually selected mean glottal flow pulses for each vocal effort level and using a spectral matching filter to correctly match the voice source spectrum to a desired style. Subjective evaluations show that the proposed DNN-based method is rated comparable to the baseline method, but avoids the manual selection of the pulses and is computationally faster than a system using a spectral matching filter.

Index Terms: Speech synthesis, deep neural network, DNN, voice source modelling, vocal effort, glottal flow

1. Introduction

Statistical parametric speech synthesis, also known as hidden Markov model (HMM) based speech synthesis [1, 2], is a popular framework for synthesising speech and a good alternative for the unit selection approach [3]. It has several benefits such as the ability to vary speaking style and speaker characteristics [4–8], small memory footprint [9, 10], and robustness [11]. However, statistical speech synthesis suffers from lower segmental speech quality compared to the unit selection systems that concatenate natural speech waveforms [3]. This degradation is thought to stem mainly from three factors: a) over-simplified vocoder techniques that are incapable of representing natural speech waveforms in detail b) acoustic modelling inaccuracy, and c) over-smoothing of the generated speech parameters [2]. This paper addresses the first factor by introducing a flexible voice source model that uses a deep neural network (DNN), with the aim of better modelling variations in the voice source signal and interaction between the source and the filter.

The modelling of the excitation signal in HMM-based speech synthesis has greatly improved since the first vocoders that used a simple impulse train excitation [12]. The quality of such simple excitation is poor due to the unnatural zero-phase character of the excitation. Mixed excitation [13] and two-band excitation [14] has greatly improved the quality by mixing periodic excitation with aperiodic noise. This mixed excitation scheme is also used in the most prevalent vocoder in speech synthesis, STRAIGHT [15, 16]. Also closed-loop training [17, 18] and parametric models of the glottal flow [19, 20] have been proposed for improving the speech quality.

Since the context-dependent characteristics of the glottal flow waveform are difficult to represent using a simple parametric voice source signal, several approaches have utilised the excitation waveform *per se* in order to preserve the natural characteristics in the waveform. The idea is not new (see e.g. [21–23]), but the development of statistical speech synthesis and vocoders have given new applications for the approach. Recently, natural glottal flow pulses or residual waveforms have been used in several vocoding approaches [24–31].

Reproducing different speaking styles has long been the strength of statistical speech synthesis. Through adaptation and similar techniques, a continuous degree of varying style can be reproduced [4–8, 32, 33]. However, only few studies have explicitly investigated the modelling of the changes in the excitation waveform in response to changes in speaking style. In consequence, while mostly changing the pitch and overall spectrum, the changes in the voice characteristics are rather limited compared to natural speech. In contrast, the experiments in [33, 34] have shown that by using an appropriate glottal flow pulse for synthesising a specific style, improvements in the perceived impression of the style are achieved. However, the current approaches need human intervention, such as manually extracting and selecting the style-specific excitation waveforms.

The aim of this work is to present and extend the work on the DNN-based voice source modelling method, preliminary presented in [35], and apply it to the reproduction of various vocal effort levels similar to the study in [33]. The new DNN-based voice source modelling method is based on learning a mapping between the acoustic features and the time-domain glottal flow waveform using DNN. Thus, in synthesis, the excitation waveform can be directly generated from the acoustic features. Subjective evaluations are performed to find out if the new simpler and automatic DNN-based method can reproduce the same quality and impression of vocal effort as the previously published method without manual intervention.

2. DNN-based voice source modelling

The proposed DNN-based voice source modelling method and its use in synthesis of various speaking styles is illustrated in Figure 1. In the training part, acoustic features are first extracted from a speech database at 5-ms intervals. As the aim is to reproduce different speaking styles, the speech database should contain both normal and style-specific speech, labelled accordingly. The feature extraction uses iterative adaptive inverse filtering (IAIF) [36] in order to decompose speech signals into the vocal tract filter and the voice source signal. This enables the further parametrisation of the voice source characteristics and the segmentation of the glottal flow waveforms. Speech features described in Table 1 are extracted, i.e., the fundamental frequency (F0), frame energy, harmonic-to-noise ratio (HNR) of five frequency bands, voice source linear prediction (LP) spectrum converted to line spectral frequencies (LSF), and vocal tract LP spectrum converted to LSF. The acoustic features of normal style are used for training an HMM-based voice, after which it can be adapted to different speaking styles.

The output voice source signal by the IAIF algorithm is used for extracting pitch-synchronous glottal flow pulse segments. First, glottal closure instants (GCIs) are detected from the differentiated glottal flow signal using peak picking at fundamental period intervals, and two-pitch-period, GCI-centred glottal flow waveform segments are extracted. The pulse segments are interpolated to a constant duration of 25 ms (400 samples at 16 kHz sampling rate), windowed with the Hanning window, and normalised in energy. The pulses are stored in a codebook and linked with the corresponding acoustics features of the frame. The duration of the pulses is selected as a compromise between minimising the amount of data stored and limiting the loss of spectral information in the pulses. A mapping between the acoustic features and the glottal flow waveform segments is established by training a DNN. Random initialisation of the DNN weights is used, after which back-propagation is applied. In order to train a flexible voice source model, speech parameters from all speaking styles were used for the DNN training.

The normal voice is adapted as in [8] to a desired style using the style-specific data and an interpolation/extrapolation coefficient, which defines the amount of adaptation from the normal voice to the desired style. After the adaptation of the voice, style-specific acoustic features are generated from context-dependent HMMs (CD-HMM) according to text input as in [25]. The acoustic features are used as input to DNN, which outputs the context and style-specific glottal flow waveforms. The generated glottal flow waveforms are interpolated to a desired length according to F0, scaled in energy, and mixed with noise according to the HNR measure as in [31]. The individual two-pitch-period waveforms are overlap-added in order to create a continuous excitation, which is filtered with the vocal tract filter generated from HMMs to create speech.

3. Experiments

3.1. Speech material

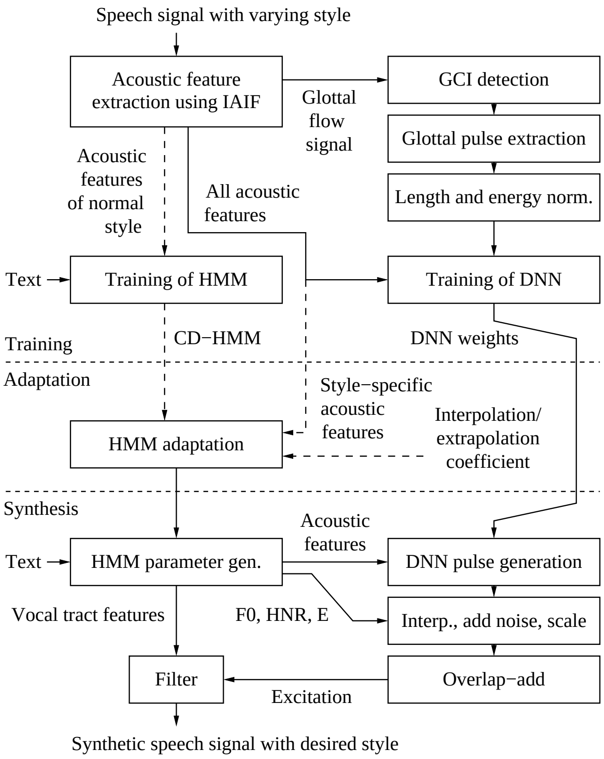

Figure 1: *Illustration of the proposed DNN-based voice source modelling method for synthesis of varying speaking styles.*

Two speech corpora, a male and a female speaker [33], were used in the experiments. For both speakers, three different vocal effort levels were utilised: breathy, normal, and Lombard. The normal style consists of 1450 sentences, comprising approximately two hours of speech for both speakers. Lombard speech was elicited by playing babble noise with 80 dB SPL to the speaker's ears through headphones while recording, and feeding back the speaker's own voice through headphones, corresponding to a level of speaking in a normal room without headphones. The Lombard style consists of 300 sentences. The breathy speaking style was elicited by increasing the level of the speaker's feedback through headphones as well as instructing the subjects to speak softly without whispering. 200 sentences were read in the breathy style. The recording and processing of the speech data are described in more detail in [33].

Table 1: *Acoustic features used for training the HMM-based voice and the DNN-based voice source model.*

Feature	Number of parameters
Energy	1
Fundamental frequency	1
Harmonic-to-noise ratio	5
Voice source spectrum	10
Vocal tract spectrum	30

3.2. Training of deep neural networks

A DNN [37] is a feed-forward, artificial neural network that has at least two layers of hidden units between input and output layers. Recently, DNNs have been successfully used for both automatic speech recognition [37] and speech synthesis [38], and DNNs have shown improvements over conventional HMM-based systems. In this work, a DNN is used in conjunction with an HMM-based approach for mapping between the acoustic features and the time-domain glottal flow waveform. The input for the DNN is the 47-dimensional acoustic feature vector, consisting of the features described in Table 1, and the output is the 400 sample duration normalised glottal flow waveform. For the hidden and output layers, sigmoid and linear activation functions are used, respectively. The DNN is trained by back-propagating derivatives of the mean squared error (MSE) cost function that measures the discrepancy between the target and actual outputs.

Previously in our research on DNN-based voice source modelling [35], a rather large DNN with two hidden layers and 1000 neurons per layer was proposed for learning the mapping between the acoustic features and the glottal flow waveform. In our recent studies, smaller DNN architectures have been shown to learn the mapping more efficiently, while also taking less

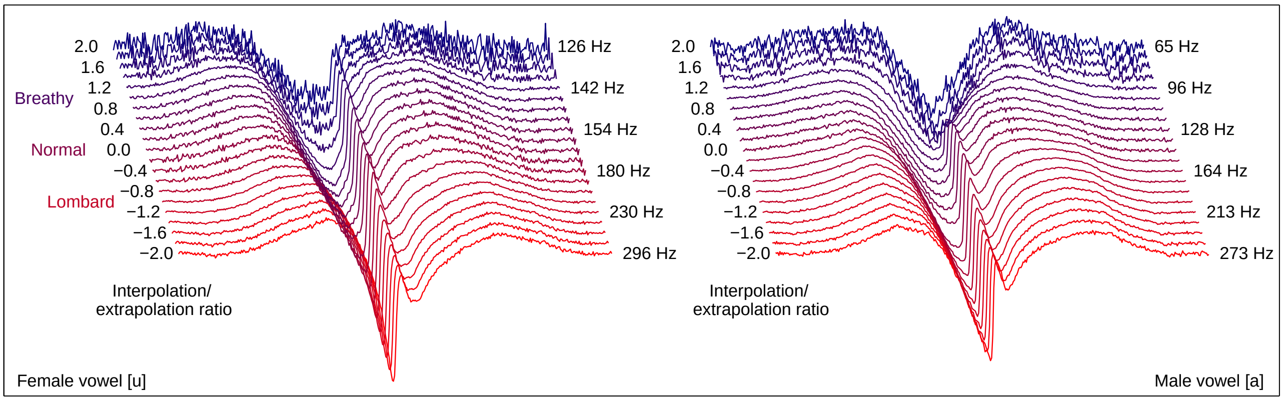

Figure 2: *Demonstration of the DNN-based excitation modelling by interpolating and extrapolating different HMM-based speaking styles from original breathy* (1.0)*, normal* (0.0)*, and Lombard speech* (−1.0)*, and generating the DNN-based pulses corresponding to the generated speech parameters of various degrees of the styles. The resulting pulses (without interpolation in time, scaling in magnitude or adding noise) are shown for female vowel [u] and male vowel [a].*

time to train. In this work, a 2-hidden-layer DNN is used with 100 and 200 neurons in the first and the second hidden layers, respectively, which achieved convergence much earlier and gave lower final errors than the DNN architecture in [35].

In addition, restricted Boltzmann machine (RBM) pre-training, used in [35], turned out to achieve fast initial reduction in training error, but the error curves saturated also very rapidly and did not achieve even nearly as low errors as random weight initialisation. This seems to indicate that the RBM pre-training helped in learning the main characteristics of the glottal flow waveform, but reduced the flexibility of the model to learn the multitude of variations in the glottal waveform shape. Thus, in this work, random initialisation of the DNN weights is used.

Since the 400-sample-length glottal flow waveform is rather high-dimensional, an approach using principal component analysis (PCA) was also experimented with. The glottal flow waveforms in the training database were decomposed into 40 principal components (PCs) and the corresponding weights, and a mapping between the 47 acoustic features and the 40 PCs were then established using a DNN. The results were similar to the sample-based approach, but glottal flow waveform was inconsistent when using unseen or noisy input data. Thus, the time-domain glottal flow waveform is used in this work.

Due to occasional small errors in the GCI estimation, and due to the averaging effect of the DNN training, the GCI peaks of the generated pulses are slightly smoother than those in the waveform inverse filtered from natural speech. In order to compensate this constant difference in spectral domain, a fixed pre-emphasis is applied at synthesis stage. The amount of pre-emphasis is estimated by comparing the spectra of the voice source signals over all styles synthesised with the DNN-based method and conventional GlottHMM synthesis using natural glottal flow pulse and a source spectral matching scheme [25]. Best match between the two spectra was achieved with first-order differentiator with $\alpha = 0.387$.

3.3. Handling data sparsity

Robustness to data sparsity is a crucial property of a generative model, and especially in speech synthesis, data sparsity is a common problem. It is often not possible to include all possible input cases in the training material, and thus it is important that a model can interpolate or extrapolate an appropriate output from input parameters that are not included in the training set.

In order to demonstrate the ability of the proposed DNN-based voice source model to create natural glottal flow waveform despite data sparsity, two training sets were constructed with the other one missing a part of the input parameter values. A data set of 280,651 input vectors and output pulse waveforms were used to train a baseline DNN using the male speech data. Since energy of the speech frame is highly dependent on the speaking style, and the glottal pulse waveform shows considerable changes in relation to changes in energy (see [35]), it was chosen as a feature to be altered in this experiment. The energy in the original training set ranged from −23.3 dB to 41.0 dB. A modified training set was constructed by removing all data points with energy values from 0 dB to 15 dB. After discarding the specific data, the number of training samples in the modified set was 227,777, removing around 19% of the total samples and corresponding exemplars of glottal flow waveforms. Both DNNs were trained similarly and the errors of the generated glottal flow waveforms were measured using a test set with i) all data, ii) in-domain data, iii) out-of-domain data. The mean, maximum, and minimum relative change in errors (E) are shown in Table 2. The results show that the overall error is only slightly increased when moving from the in-domain data (0.73%) to the out-of-domain data (2.07%), indicating that the model can rather successfully interpolate/extrapolate the output.

3.4. Voice building

The HMM training and adaptation procedures were identical to the experiments done in [33]. The training of the normal voices followed the standard HTS method [39]. Speech features described in Table 1 were extracted using the GlottHMM vocoder [25] and delta, and delta-delta features were added. Semi hidden Markov models were used as acoustic models, and features were trained in individual streams except the vocal tract LSFs and energy were trained together.

Table 2: *Mean, maximum, and minimum change in the error E over the generated glottal flow waveforms when using a training data with induced data sparsity in comparison to using all data.*

Test data	Δmean(E)	Δmax(E)	Δmin(E)
All data	1.17 %	1.37 %	−7.86 %
In-domain data	0.73 %	1.37 %	−7.86 %
Out-of-domain data	2.07 %	−1.89 %	36.18 %

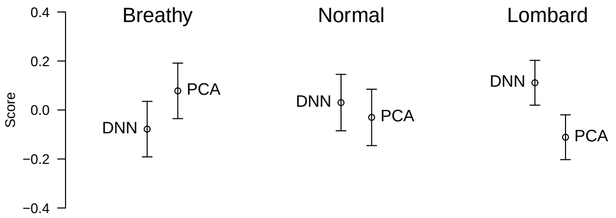

Figure 3: *Results of the quality test.*

In order to create the low and high vocal effort voices (breathy and Lombard), the normal voice models were adapted with constrained structural maximum a posteriori linear regression combined with maximum a posteriori (CSMAPLR + MAP) adaptation technique [8]. The speaker-dependent voice source model DNNs were trained using all speech material including breathy, normal, and Lombard speech.

For demonstrating the interpolation and extrapolation characteristics of the DNN-based voice source modelling, both voices were adapted to various degrees of vocal effort from very breathy to very Lombard, and corresponding speech parameters were generated. Normal training voices being at point 0.0 and adaptation samples at points 1.0 (breathy) and at -1.0 (Lombard), adapted voices were created between 2.0 (very breathy) and -2.0 (very Lombard) with a step size of 0.2. The parameters of each voice were then fed to the DNNs to generate style-specific glottal flow waveforms. Generated waveforms for female vowel [u] and male vowel [a] are shown in Figure 2.

3.5. Subjective evaluation

In order to evaluate the performance of the proposed method, subjective evaluations were conducted using three vocal effort levels. The final voices used in the subjective evaluation were created at points 1.0 (breathy), 0.0 (normal), and -1.0 (Lombard) for both speakers. A high-quality mean glottal flow pulse excitation scheme was selected for a reference baseline system, which has been successfully used in synthesising speech with varying vocal effort [33]. The baseline system uses a style-specific mean glottal flow pulse for each of the three styles [33] (corresponding to the PCA-based excitation in [29]), and a spectral matching scheme [25,33], where a pole-zero filter is used to filter the excitation signal in order to apply the desired spectral properties defined by the generated voice source spectrum.

Two types of tests were conducted to compare the proposed and the baseline systems. First, a comparison category rating (CCR) test was conducted to evaluate the speech quality. In a CCR test, listener hears two different samples and rates the quality difference between them on the 7-point comparison mean opinion score scale ranging from much worse (-3) to much better (3). A total of 14 native Finnish listeners evaluated 60 sample pairs each, and the preference of the methods was evaluated by averaging the listener scores for each method.

Secondly, a similarity test was conducted in order to assess the speaker and style similarity between the two methods. In the similarity test, listener is presented with two speech samples synthesised by the two methods, and a natural reference sample corresponding to the speaker and style of the synthetic samples. The task of the listener is to choose which of the two samples is more similar to the reference in terms of speaker and style, or no preference between the samples. A total of 14 native Finnish listeners evaluated 60 sample pairs each.

The mean scores of the quality test are shown for each vocal effort level with 95% confidence intervals in Figure 3. Only with Lombard speech, the difference between the two methods is statistically significant with the proposed method being rated higher in quality. Figure 3 presents the results of the similarity test, showing the proportion of answers (with 95% confidence intervals) for each method and for each vocal effort level. Only in the case of breathy speech, the results are statistically significant, where the baseline method is rated more similar.

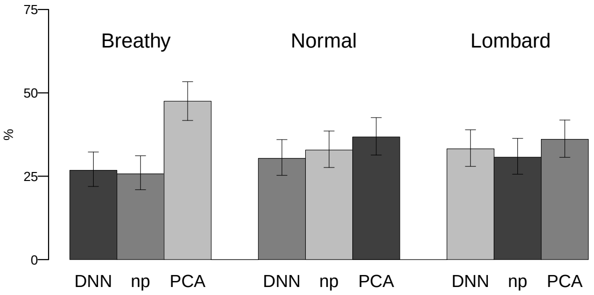

Figure 4: *Results of the similarity test. The middle bars labelled as* np *stand for no preference for either of the methods.*

4. Discussion and conclusions

The experiments show that the proposed DNN-based voice source modelling method is capable of successfully reproducing different degrees of vocal effort, and that it improves the synthesis quality with Lombard speech in comparison to the baseline method. Although the proposed method was able to generate breathier waveforms than the baseline system, and although the resulting breathy voice was perceptually softer based on informal listener reports, the similarity of the breathy voice was slightly decreased due to the absence of the spectral matching, as is used in [25,33]. In comparison to the baseline method, the proposed method avoids manual intervention needed for the voice style variation, and enables continuous style variation within an utterance, which is required for plausible expressive speech synthesis. Moreover, the generation of pulses from a DNN is computationally less expensive than filtering the excitation signal with a pole-zero spectral matching filter.

The study shows that the approximate shape of the glottal flow waveform can be successfully modelled by the proposed approach in order to generate various speaking styles. However, the proposed method does not seem to greatly improve the segmental quality of speech compared to using a pre-selected glottal flow pulses. This indicates that even though the DNN-based modelling approach is capable of generating the gross shape of the glottal flow pulse, it introduces an averaging effect that removes detailed variations of the pulse needed to achieve quality close to natural speech.

The existence of interaction between the source and filter is well known (see e.g. [40]), but it is hardly utilised in speech technology or in speech synthesis. In this work, the glottal flow waveform is predicted based on features including the vocal tract spectrum, but it seems that modelling the source and filter interaction by the proposed method is not adequate for greatly improving the segmental speech quality. Future directions of the study will be concentrated on the more accurate modelling of the source-filter interaction using the DNN-based approach.

5. Acknowledgements

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

6. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis,” in *Proc. Eurospeech*, 1999, pp. 2374–2350.

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

[3] A. Hunt and A. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, 1996, pp. 373–376.

[4] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, “Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR,” in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, 2001, pp. 805–808.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Speaker interpolation in HMM-based speech synthesis system,” in *Proc. Eurospeech*, 1997, pp. 2523–2526.

[6] K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Eigenvoices for HMM-based speech synthesis,” in *Proc. ICSLP*, 2002, pp. 1269–1272.

[7] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, “A style control technique for HMM-based expressive speech synthesis,” *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 9, pp. 1406–1413, 2007.

[8] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, “Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm,” *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 17, no. 1, pp. 66–83, 2009.

[9] S.-J. Kim, J.-J. Kim, and M.-S. Hahn, “HMM-based Korean speech synthesis system for hand-held devices,” *IEEE Trans. Consum. Electron.*, vol. 52, no. 4, pp. 1384–1390, 2006.

[10] A. Gutkin, X. Gonzalvo, S. Breuer, and P. Taylor, “Quantized HMMs for low footprint text-to-speech synthesis,” in *Proc. Interspeech*, 2010, pp. 837–840.

[11] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, “Robust speaker-adaptive HMM-based text-to-speech synthesis,” *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 17, no. 6, pp. 1208–1230, 2009.

[12] J. Makhoul, “Linear prediction: A tutorial review,” *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[13] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Mixed excitation for HMM-based speech synthesis,” in *Proc. Eurospeech*, 2001, pp. 2259–2262.

[14] S. J. Kim and M. Hahn, “Two-band excitation for HMM-based speech synthesis,” *IEICE Trans. Inf. & Syst.*, vol. E90-D, 2007.

[15] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, “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 Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999.

[16] 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 *2nd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, 2001.

[17] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda, “An excitation model for HMM-based speech synthesis based on residual modeling,” in *6th ISCA Speech Synthesis Workshop*, 2007.

[18] R. Maia, H. Zen, and M. J. F. Gales, “Statistical parametric speech synthesis with joint estimation of acoustic and excitation model parameters,” in *7th ISCA Speech Synthesis Workshop*, 2010, pp. 88–93.

[19] J. Cabral, S. Renalds, K. Richmond, and J. Yamagishi, “Towards an improved modeling of the glottal source in statistical parametric speech synthesis,” in *Sixth ISCA Workshop on Speech Synthesis*, 2007, pp. 113–118.

[20] ——, “Glottal spectral separation for parametric speech synthesis,” in *Proc. Interspeech*, 2008, pp. 1829–1832.

[21] J. Holmes, “The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer,” *IEEE Trans. Audio and Electroacoustics*, vol. 21, no. 3, pp. 298–305, 1973.

[22] K. Matsui, S. D. Pearson, K. Hata, and T. Kamai, “Improving naturalness in text-to-speech synthesis using natural glottal source,” in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, vol. 2, 1991, pp. 769–772.

[23] G. Fries, “Hybrid time- and frequency-domain speech synthesis with extended glottal source generation,” in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, vol. 1, 1994, pp. 581–584.

[24] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, “HMM-based Finnish text-to-speech system utilizing glottal inverse filtering,” in *Proc. Interspeech*, 2008, pp. 1881–1884.

[25] 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. Audio, Speech and Lang. Proc.*, vol. 19, no. 1, pp. 153–165, 2011.

[26] T. Drugman, G. Wilfart, A. Moinet, and T. Dutoit, “Using a pitch-synchronous residual codebook for hybrid HMM/frame selection speech synthesis,” in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, 2009, pp. 3793–3796.

[27] J. Sung, D. Hong, K. Oh, and N. Kim, “Excitation modeling based on waveform interpolation for HMM-based speech synthesis,” in *Proc. Interspeech*, 2010, pp. 813–816.

[28] T. Drugman and T. Dutoit, “The deterministic plus stochastic model of the residual signal and its applications,” *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 20, no. 3, pp. 968–981, 2012.

[29] T. Raitio, A. Suni, M. Vainio, and P. Alku, “Comparing glottal-flow-excited statistical parametric speech synthesis methods,” in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, 2013, pp. 7830–7834.

[30] T. Drugman, G. Wilfart, and T. Dutoit, “A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis,” in *Proc. Interspeech*, 2009, pp. 1779–1782.

[31] 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 Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, 2011, pp. 4564–4567.

[32] B. Picart, T. Drugman, and T. Dutoit, “Analysis and HMM-based synthesis of hypo and hyperarticulated speech,” *Computer Speech & Language*, vol. 28, no. 2, pp. 687–707, 2014.

[33] T. Raitio, A. Suni, M. Vainio, and P. Alku, “Synthesis and perception of breathy, normal, and lombard speech in the presence of noise,” *Computer Speech & Language*, vol. 28, no. 2, pp. 648–664, 2014.

[34] T. Raitio, J. Kane, T. Drugman, and C. Gobl, “HMM-based synthesis of creaky voice,” in *Proc. Interspeech*, 2013, pp. 2316–2320.

[35] 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, accepted.

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

[37] 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.

[38] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, 2013, pp. 7962–7966.

[39] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, “The HMM-based speech synthesis system (HTS) version 2.0,” in *Sixth ISCA Workshop on Speech Synthesis*, 2007, pp. 294–299.

[40] I. R. Titze, “Nonlinear source-filter coupling in phonation: Theory,” *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 2733–2749, 2008.