VOICE SOURCE MODELLING USING DEEP NEURAL NETWORKS FOR STATISTICAL PARAMETRIC SPEECH SYNTHESIS

Tuomo Raitio[*], *Heng Lu*[†], *John Kane*[‡], *Antti Suni*[§], *Martti Vainio*[§], *Simon King*[†], *Paavo Alku*[*]

[*] Department of Signal Processing and Acoustics, Aalto University, Finland
[†] Centre for Speech Technology Research, University of Edinburgh, UK
[‡] Phonetics and Speech Laboratory, Trinity College Dublin, Ireland
[§] Institute of Behavioural Sciences, University of Helsinki, Finland

ABSTRACT

This paper presents a voice source modelling method employing a deep neural network (DNN) to map from acoustic features to the time-domain glottal flow waveform. First, acoustic features and the glottal flow signal are estimated from each frame of the speech database. Pitch-synchronous glottal flow time-domain waveforms are extracted, interpolated to a constant duration, and stored in a codebook. Then, a DNN is trained to map from acoustic features to these duration-normalised glottal waveforms. At synthesis time, acoustic features are generated from a statistical parametric model, and from these, the trained DNN predicts the glottal flow waveform. Illustrations are provided to demonstrate that the proposed method successfully synthesises the glottal flow waveform and enables easy modification of the waveform by adjusting the input values to the DNN. In a subjective listening test, the proposed method was rated as equal to a high-quality method employing a stored glottal flow waveform.

Index Terms— Deep neural network, DNN, voice source modelling, glottal flow, statistical parametric speech synthesis

1. INTRODUCTION

Statistical parametric speech synthesis, or hidden Markov model (HMM) speech synthesis [1,2], is a flexible framework for synthesising speech. It has several attractive properties, such as the ability to vary speaking style and speaker characteristics, small memory footprint, and robustness. However, HMM-based speech synthesis suffers from lower speech quality than the unit selection approach [3] and this is thought to stem mainly from three factors: a) over-simplified vocoder techniques, b) acoustic modelling inaccuracy, and c) over-smoothing of the generated speech parameters [2]. This paper addresses the problem of over-simplified vocoders by introducing a new voice source modelling method using a deep neural network (DNN).

One of the key factors in improving the quality of statistical speech synthesis has been the development of better excitation modelling techniques. The earliest vocoders used a train of impulses [4] located at the glottal closure instants to model voiced excitation. The quality of this impulse-train-excited speech is poor with a buzzy sound quality due to the zero-phase character of the excitation. Several improvements, such as mixed excitation [5] and two-band excitation [6], have been introduced to alleviate this effect by mixing periodic excitation with aperiodic noise. Mixed excitation is used in, e.g., STRAIGHT [7, 8], which is one of the most widely used vocoders in HMM-based speech synthesis. Voiced excitation has also been improved by using a closed-loop training approach [9,10] or parametric models of the glottal flow [11,12].

The natural excitation of voiced speech, the glottal flow, is difficult to represent as a compressed parametric vector suitable for statistical parametric modelling. Therefore, sampling approaches that utilise the excitation waveform *per se* have been proposed that capture the detailed characteristics of the signal. This idea is not new (see e.g. [13–15]), but the development of statistical parametric synthesis has given rise to several novel excitation methods based on natural speech samples. For example, in [16, 17], a glottal flow pulse estimated from natural speech (using glottal inverse filtering) is manipulated in order to construct a more natural excitation signal. In [18–21], principal component analysis (PCA) is applied to pitch-synchronous residual/glottal flow signals to represent the excitation waveform. In [22, 23], a pitch-synchronous residual/glottal flow codebook is constructed, from which appropriate pulses are selected for synthesis.

Yet, sampling in the voice source domain exhibits some challenges similar to those in the unit selection approach [21, 23], i.e., finding the best sequence of units that well matches the given target specification and concatenate imperceptibly together. Purely sampling-based approaches are, like unit selection, inherently inflexible and limited by the available samples in the database: this limits the ability of the system to

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287678 (Simple[4]All), the Academy of Finland, and EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

change voice quality in a continuous manner, for example.

To overcome the above problems of using stored samples without attempting to construct a fully parametric model of glottal pulses (which has proved very challenging), we introduce a novel voice source modelling technique that can be considered as a compromise between waveform sampling and parametric modelling. The method is based on predicting the pitch-synchronous glottal flow directly in the time-domain by using a DNN. The DNN is used to map the modelled speech parameters to the actual excitation waveform, which can then be used directly for synthesis in combination with predicted vocal tract features. The proposed method has the flexibility of a parametric model because it is able to generate variation in the voice source waveform in response to changes in the speech features. It also exhibits some of the advantages of stored sample-based methods in that the predicted waveforms contain more detail than parametric models.

The paper is organised as follows. First, DNNs in the context of this work are introduced in Sec. 2, after which the proposed DNN-based voice source modelling technique is described in Sec. 3. Experiments using the new method are described in Sec. 4, concentrating on DNN architecture and training, and on the use of the proposed method in copy-synthesis, voice source modification, and HMM-based synthesis. Finally, Sec. 5 concludes the paper.

2. DEEP NEURAL NETWORKS

A DNN [24] is a feed-forward, artificial neural network that has at least two layers of hidden units between input and output layers. In this work, a DNN is used to build a mapping from extracted acoustic speech features to corresponding glottal flow pulses. This is a regression problem, where we are predicting continuously-valued outputs, so we chose a linear activation function for the output (regression) layer and sigmoid activation function units for the hidden layers. The latter is defined as

$$v_i = f(\sum_j W_{ij} x_j + b_i), \tag{1}$$

where $f(x) = 1/(1 + \exp(-x))$ is the sigmoid logistic function, W_{ij} and b_i are weights and biases, and x_j and v_i are the input and output of the DNN, respectively. For the linear layer, the activation function is simply

$$v_i = \sum_j W_{ij} x_j + b_i. \tag{2}$$

Restricted Boltzmann machine (RBM) pre-training can be used to prevent over-fitting to the data, which aims at unsupervised learning of the distributions of the input features. Since the input acoustic features are real valued in this work, a Gaussian–Bernoulli RBM [24] is employed for the visible (input) layer. After optional pre-training, the DNN is trained ("fine-tuned") by back-propagating derivatives of a cost function that measures the discrepancy between the target outputs and the actual outputs. In this work, mean squared error (MSE) is used as the cost function. The error function is

$$E = \sum_j (\boldsymbol{v}_j - \hat{\boldsymbol{v}}_j)^2, \tag{3}$$

where $\hat{\boldsymbol{v}}_j$ is the regression target for DNN training.

3. DNN-BASED VOICE SOURCE MODELLING

Recently, for both automatic speech recognition [24] and speech synthesis [25], DNNs have shown improvements over conventional HMMs. In this exploratory work, a DNN is used in conjunction with a HMM-based system. The approach is illustrated in Fig. 1. First, frame-wise acoustic features are extracted from a database. In the feature extraction, iterative adaptive inverse filtering (IAIF) [26] is used to decompose the speech signal into a vocal tract filter and a voice source signal. The extracted speech parameters include the vocal tract linear prediction (LP) filter that is converted to a line spectral frequency (LSF) representation, and parameters describing the properties of the voice source, i.e., fundamental frequency (F0), frame energy, harmonic-to-noise ratio (HNR) of five frequency bands, and voice source LP spectrum converted to LSF. The extracted features, depicted in Table 1, are then used to train a HMM-based synthesiser, as in [17].

The IAIF method produces an estimate of the voice source signal from which individual glottal flow pulses are extracted. To do this, glottal closure instants (GCIs) are detected from the differentiated glottal flow signal using a simple peak picking algorithm. This enables the extraction of two-pitch-period, GCI-centred glottal flow pulses, delimited by two other GCIs. The pulse segments are interpolated to a constant duration of 25 ms (400 samples at 16 kHz sampling rate), windowed with the Hann window, normalised in energy, and stored in a codebook. The fixed duration of the pulses is chosen as a compromise between minimising the amount of data stored and limiting loss of spectral information.

Given the set of glottal pulses and corresponding vectors of 47 acoustic parameters (Table 1), a mapping is established by training the DNN. RBM pre-training is used to alleviate over-fitting, after which back-propagation is applied. For synthesis, both vocal tract and voice source parameters are generated from context-dependent HMMs, as in [17]. Instead of using the source speech parameters to select a sequence of stored pulse waveforms drawn from the codebook, we use the complete set of 47 acoustic parameters as input to the DNN, which outputs glottal flow derivative waveforms. The generated glottal flow pulses are interpolated to a duration corresponding to the required F0, scaled in energy, mixed with noise according to the HNR measure, and overlap-added to generate the excitation for synthesis. Alternatively, the DNN

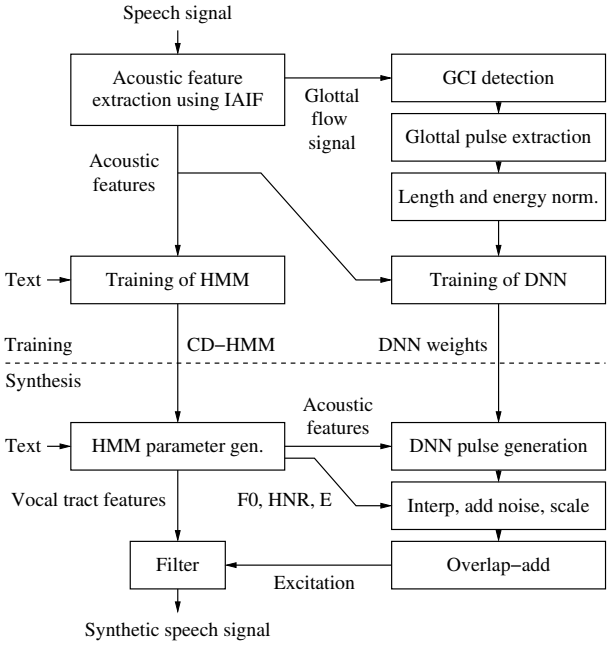

Fig. 1. Illustration of the proposed HMM-based speech synthesis using DNN-based voice source modelling.

pulses can be used as a target for selecting the closest matching stored glottal flow waveforms from the codebook (similar to [23]). The latter method has two potential benefits: 1) the natural codebook pulses preserve the detailed source waveform, and 2) the DNN target pulse prevents the selection of spurious pulses from the codebook. The vocal tract filter already generated by the HMM is then used to filter the excitation signal, producing synthetic speech.

4. EXPERIMENTS

4.1. Experimental setup

Two Finnish speech databases, male *MV* and female *Heini*, recorded for the purpose of speech synthesis, were used in the experiments. The male voice comprises 600 sentences (approx. 1 h of speech) and the female database comprises 500 sentences. Both voices were sampled at 16 kHz.

The GlottHMM vocoder [17, 23] was used for extracting the acoustic features and the glottal flow signal using IAIF. Glottal flow pulse codebooks were constructed for both databases in order to train the DNN-based voice source model. The codebooks contained 203,172 and 203,768 pulses for the male and female speakers, respectively. Additionally, smaller codebooks were constructed for both speakers from 20 sentences of speech material, in order to implement the alternative method in which the DNN output is used to select a natural pulse from the codebook; these codebooks consisted only of 7,495 and 8,131 pulses in order to minimise computational cost at synthesis time. Previous experiments [23] have shown that using a much larger codebook does not significantly improve the synthesis quality. The standard HTS 2.1 method [27] was used for training the HMM-based system.

Table 1. Acoustic features used for training the HMM-based synthesis 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

4.2. DNN training

The DNN as described in Sec. 2 is used. The input is the 47-dimensional vector composed of the extracted acoustic speech features listed in Table 1 and the target output is a 400 sample duration normalised glottal flow pulse.

In order to determine the optimal number of layers and hidden units for DNN, six different systems (A–F) were trained by varying the number of hidden layers (from 1 to 3) and the number of units per layer (from 800 to 1200). Unsupervised RBM pre-training was tried for one configuration. 200,000 training examples were used for training with 3,000 examples for cross-validation. The training and development errors for each system are presented in Table 2. The results show that system F with 2 hidden layers and 1000 units per hidden layer gave best results, with RBM pre-training slightly improving performance (compare system F to system B).

4.3. Voice source modelling and modification

Copy-synthesis for unseen speech data (i.e., not in the training or validation sets) using the proposed method is illustrated in Fig. 3, which shows the original (differentiated) excitation estimated by IAIF from natural speech and the synthetic DNN-based excitation generated from the extracted parameters. In informal listening, the proposed voice source modelling method produces natural sounding copy-synthesis, either by directly using the DNN generated pulses or by using them as a target to select pulses from the smaller codebook.

The advantage of predicting pulses with the DNN is the ability to continuously adjust the glottal flow waveform in response to the input acoustic features. Fig. 2 demonstrates this ability (see last page): frame energy, F0, and HNR are varied

	Hidden layers	Units per layer	Pre-training	Train error	Dev set error
A	1	1000	No	0.411	0.499
B	2	1000	No	0.398	0.488
C	3	1000	No	0.400	0.489
D	2	800	No	0.404	0.493
E	2	1200	No	0.413	0.502
F	2	1000	Yes	0.394	0.485

Table 2. Training and development mean squared error (MSE) for various DNN configurations.

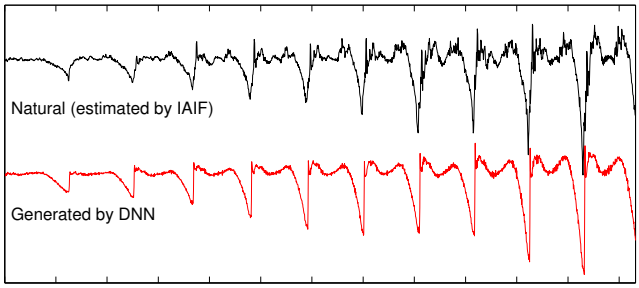

Fig. 3. Demonstration of the DNN-based excitation generation by copy-synthesis of a Finnish male speech segment [vie]. The upper signal (black) is the differentiated glottal flow estimated by IAIF. The lower signal (red) is the excitation generated by DNN according to the extracted features with noise mixed in according to HNR.

individually while other parameters are left unchanged, and pulses are generated from the trained DNN. The pulse waveform displays a continuous and consistent change in response to the varied speech parameter. For example, with low input energy, the glottal pulse shows a less prominent peak at the GCI whilst with high input energy the pulse has a very sharp discontinuity at the GCI. Similarly natural behaviour is observed also with F0 and HNR. This opens up possibilities for more flexible voice source modification.

4.4. Subjective evaluation of HMM synthesis

In order to demonstrate the capability and assess the quality of the proposed method, an online subjective evaluation was carried out. Three different methods were chosen for comparison: 1) Conventional GlottHMM synthesis [17] using a single natural glottal flow pulse of which spectrum is matched according to the voice source LSF (*Pulse*), 2) DNN-based voice source modelling (*DNN*), and 3) DNN-based voice source model used as a target cost for selecting pulses from a codebook (*DNN-c*). The latest single pulse GlottHMM was selected for comparison since it has been found to be a reliable method for producing high quality synthetic speech, and better than STRAIGHT with male speech [17]. Thus, the baseline method can be considered to represent state-of-the-art.

A comparison category rating (CCR) test was used, in which pairs of stimuli are presented to participants, whose task is to indicate the difference between the two samples on a seven-point CMOS scale ranging from much worse (-3) to much better (3). All three combinations of the systems (1–2, 1–3, 2–3) were evaluated. 50 utterances were synthesised from held-out data from both speakers and for each of the three systems (300 stimuli in total). In order to reduce the workload on participants, 10 sentences from both speakers were randomly selected for each participant and presented to them in each of the three system combinations. Thus each participant rated a total of 60 stimuli pairs. Also the ordering of the pairs of stimuli was randomised. 26 people (15 Finnish and 11 non-Finnish) participated in the evaluation. The CCR

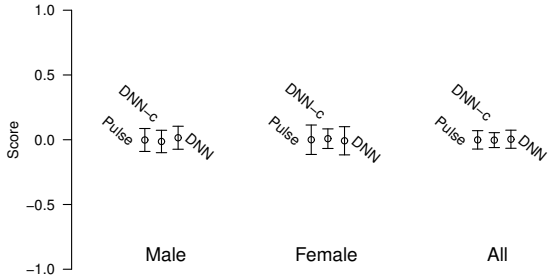

Fig. 4. Results of the subjective evaluation.

test responses are summarised by calculating the mean score for each method, which yields the order of preference and distances between all the methods (i.e., the amount of preference relative to each other). The results of the CCR test, plotted in Fig. 4, are encouraging in showing that both DNN-based methods are rated as equal to the high-quality baseline system. The differences in quality between the compared systems are rather small due to the read-aloud voice quality. With more expressive speech material the proposed methods are expected to provide more advantage over the baseline.

5. CONCLUSIONS

This paper presented a voice source modelling method based on predicting the time domain glottal flow waveform using a DNN. In the experiments presented in this paper, the proposed DNN-based method is shown to successfully generate acoustic feature-dependent glottal flow waveforms and to produce high-quality HMM-synthesis, comparable to the state-of-the-art methods. In addition to accurate voice source generation in synthesis, the method offers possibilities for automatic or manual voice source modification. In future work, the proposed method will be assessed using more expressive speech material where the new method is expected to show more advantage over conventional methods. Also the mapping from the acoustic features to the glottal flow waveform will be further studied by exploring different DNN architectures.

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] Heiga Zen, Keiichi Tokuda, and Alan 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. Acoust. Speech Signal Proc.*, 1996, pp. 373–376.

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

[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.

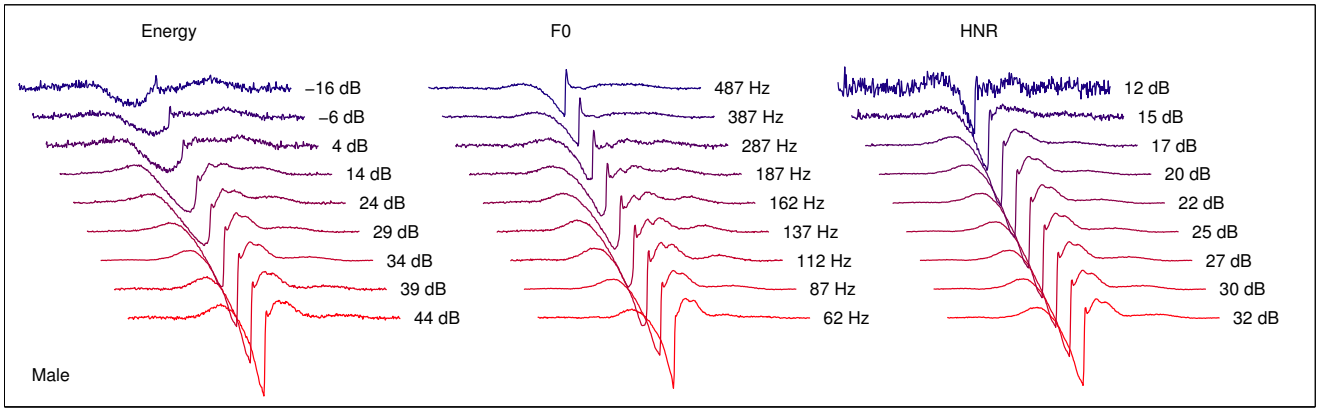

Fig. 2. Demonstration of the DNN-based excitation generation by adjusting the input parameters to produce various different pulses. Energy, F0, and HNR are adjusted within and slightly over the values present in the original training data. The resulting pulses (without interpolation, scaling or adding noise) are shown for male vowel [i] of normal phonation for each of the three adjusted parameters and corresponding values. During the adjustment of one parameter, others were kept constant.

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

[7] 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.

[8] H. Kawahara, Jo 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.

[9] 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 Workshop on Speech Synthesis*, 2007.

[10] 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.

[11] J. Cabral, S. Renalds, K. Richmond, and J. Yamagishi, "Towards an improved modeling of the glottal source in statistical parametric speech synthesis," in *6th ISCA Workshop on Speech Synthesis*, 2007, pp. 113–118.

[12] J. Cabral, S. Renalds, K. Richmond, and J. Yamagishi, "Glottal spectral separation for parametric speech synthesis," in *Proc. Interspeech*, 2008, pp. 1829–1832.

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

[14] 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. Acoust. Speech Signal Proc.*, 1991, vol. 2, pp. 769–772.

[15] G. Fries, "Hybrid time- and frequency-domain speech synthesis with extended glottal source generation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Proc.*, 1994, vol. 1, pp. 581–584.

[16] 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.

[17] 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 Lang. Proc.*, vol. 19, no. 1, pp. 153–165, 2011.

[18] 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.

[19] 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.

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

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

[22] 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. Acoust. Speech Signal Proc.*, 2009, pp. 3793–3796.

[23] 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. Acoust. Speech Signal Proc.*, 2011, pp. 4564–4567.

[24] 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.

[25] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Proc.*, 2013, pp. 7962–7966.

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

[27] 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 *6th ISCA Workshop on Speech Synthesis*, 2007, pp. 294–299.