Multimedia Tools and Applications
https://doi.org/10.1007/s11042-020-09783-9

Noise and acoustic modeling with waveform generator in text-to-speech and neutral speech conversion

Mohammed Salah Al-Radhi[1] · Tamás Gábor Csapó[1,2] · Géza Németh[1]

Received: 14 April 2020 / Revised: 31 July 2020 / Accepted: 28 August 2020
Published online: 10 September 2020

Abstract

This article focuses on developing a system for high-quality synthesized and converted speech by addressing three fundamental principles. Although the noise-like component in the state-of-the-art parametric vocoders (for example, STRAIGHT) is often not accurate enough, a novel analytical approach for modeling unvoiced excitations using a temporal envelope is proposed. Discrete All Pole, Frequency Domain Linear Prediction, Low Pass Filter, and True envelopes are firstly studied and applied to the noise excitation signal in our continuous vocoder. Second, we build a deep learning model based text–to–speech (TTS) which converts written text into human-like speech with a feed-forward and several sequence-to-sequence models (long short-term memory, gated recurrent unit, and hybrid model). Third, a new voice conversion system is proposed using a continuous fundamental frequency to provide accurate time-aligned voiced segments. The results have been evaluated in terms of objective measures and subjective listening tests. Experimental results showed that the proposed models achieved the highest speaker similarity and better quality compared with the other conventional methods.

Keywords Speech synthesis · Vocoder · Temporal envelope · Neural network · Voice conversion

✉ Mohammed Salah Al-Radhi
malradhi@tmit.bme.hu

Tamás Gábor Csapó
csapot@tmit.bme.hu

Géza Németh
nemeth@tmit.bme.hu

[1] Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary

[2] MTA-ELTE Lendület Lingual Articulation Research Group, Budapest, Hungary

1 Introduction

Speech synthesis can be defined as the ability to produce human speech by a machine like computer. Statistical parametric speech synthesis (SPSS) using waveform parametrisation has recently attracted much interest caused by the advancement of Hidden Markov Model (HMM) [62] and deep neural network (DNN) [63] based text-to-speech (TTS). Such a statistical framework that can be guided by an analysis/synthesis system (which is also called vocoder) is used to generate human voice from mathematical models of the vocal tract. Although there are several different types of vocoders (e.g., see [43] for comparison) that use analysis/synthesis, they follow the same main strategy. During the analysis phase, vocoder parameters are extracted from the speech waveform which represent the excitation speech signal and filter transfer function (spectral envelope). On the other hand, in the synthesis phase, the vocoded parameters are interpolated over the current frame across the synthesis filter to reconstruct the speech signal. Since the design of a vocoder depends on speech characteristics, the quality of synthesized speech may still be unsatisfactory. Thus, in this research, we aim to develop a more flexible and innovative vocoder-based speech synthesis.

In voiced sounds, it is necessary to determine the rate of vibration of the vocal cords, called fundamental frequency (F0 or pitch). Note that F0 values are continuous in the voiced parts and discontinuous in the unvoiced parts. Therefore, modeling these parts accurately will be complicated. Multi-Space Probability Distribution (MSPD) was suggested in [57] and frequently accepted, but not optimal [24] due to the discontinuities between voiced and unvoiced segments. Among other methods used to solve this problem, continuous F0 (contF0) was proposed in [25] based on HMM, which means that contF0 observations are also expected to be in unvoiced parts. Moreover, a simple contF0 estimation algorithm was found in [19] to be more productive, more accurate, and less complicated in achieving natural synthesized speech. Maximum Voiced Frequency (MVF) is another time-varying parameter that leads to a high quality of synthetic speech [12]. Unlike voicing decisions in traditional vocoders with discontinuous F0, MVF can be employed as a boundary between the periodic and aperiodic bands in the frequency domain.

An advantageous approach to reconstruct the features of voiced frames is that of the time-domain envelope, which previously introduced in speech intelligibility [13]. There are several different techniques proposed and adequately addressed in the literature to achieve a more accurate representation of temporal envelopes [3, 42, 46, 49, 51], each with their strengths and weakness. In SPSS, the envelope-based approach is successfully used to improve the quality of a source model [5, 11, 33]; however, their parameters may need further adjustment and customization in vocoding. Therefore, this paper studies a new method, which does not require any further tuning adjustment.

Over the years, vocoders have mostly motivated on modeling the different kinds of speech utterances within the same model. For example, coexistence of periodic and aperiodic components is proposed in [12], and a uniform phase used to represent voiced and unvoiced speech segments [8]. The most commonly used vocoder STRAIGHT [27] was proposed as an efficient model-based SPSS to achieve near-natural speech synthesis. In the analysis phase, STRAIGHT decomposes the speech signal into fundamental frequency, band-aperiodicity, and spectrogram parameters. On the other hand, in the synthesis phase, STRAIGHT uses a mixed approach between voiced and unvoiced excitations in a speech signal. However, it is computationally intensive and hard to meet the real-time requirement. Although some vocoder-based methods have been recently developed and applied to produce high-quality speech synthesis, the noise component is still not correctly modelled even in the most sophisticated STRAIGHT vocoder. Some experiments have attempted to overcome this problem by using, such as noise-masking approach [8] or amplitude time envelope of aperiodic components [15]. Therefore, the first goal in this paper is to shape the higher frequency components of the excitation by approximating the residual temporal envelope to provide a high accurate approximation to the actual speech signal.

In recent times, deep learning and neural networks have become the most common types of acoustic models used in SPSS for obtaining high-level data representations [37], that a significant improvement in speech quality can be achieved. In this approach, feed-forward deep neural networks (FF-DNNs) that exploit many layers of nonlinear information processing are introduced to deal with the high-dimensional acoustic parameters [10], multi-task learning [59], and replace the decision tree used in HMM systems [63]. Even though FF-DNNs were successfully applied to the acoustic modeling of speech synthesis, the high learning cost with multiple hidden layers is still a critical bottleneck [61]. Furthermore, sequence modeling with arbitrary length in the FF-DNNs architecture is also ignored. To deal with these limitations, recurrent neural networks (RNNs) have an advantage in sequence-to-sequence modeling and capturing long-term dependencies [4]. The long short-term memory (LSTM) [21], bi-directional LSTM [17], and gated recurrent unit (GRU) [31] are commonly used recurrent neurons in different research fields (e.g., image and speech processing).

Consequently, the second goal of this paper is to build a deep learning-based acoustic model for speech synthesis using feedforward and recurrent neural network as an alternative to HMMs. Here, the objective is two-fold: (a) to overcome the limitation of HMMs which typically create over-smoothing and muffled synthesized speech, (b) to ensure that all continuous parameters used by the proposed vocoder were obtained during training that could synthesize very high-quality speech.

The main objective of voice conversion (VC) is to transform the characteristics of the source speaker into that of the target speaker. It has great potential in the development of various speech tasks, for instance, speaking assistance [38] and speech enhancement [55]. Numerous statistical approaches have been employed for mapping the source features (e.g. F0) to the target domain. Gaussian mixture model (GMM) [52] is a typical form of VC that requires a source-target alignment for training the conversion models. Many other methods have also been proposed and achieved improvements in speech conversion; for example, non-negative matrix factorization [60], restricted Boltzmann [40], auto-encoders [47], and maximum likelihood [54]. However, the phenomenon of the over-smoothing and the discontinuity issues of F0 made the converted speech sounds muffled, which degrades the similarity performance. In recent times, deep learning is one of the most significant advances in voice conversion and getting quite a lot of attention from researchers. Deep belief networks [39], generative adversarial networks [26], bidirectional long short-term memory [53] have been recently proposed to preserve the sound quality. Nevertheless, the similarity of the converted voices is still degraded in terms of subjective quality due to model complexity and computational expense. Moreover, modeling of discontinuous fundamental frequency in speech conversion is problematic because the voiced and unvoiced speech regions of the source speaker are typically not appropriately aligned with the target speaker. To overcome these limitations, a new model is developed to achieve more natural converted speech.

2 Proposed speech model

2.1 Vocoder description: An overview

To construct our statistical proposed approaches, a continuous vocoder [6] has been used and evaluated. It aims to overcome the discontinuity issues in the speech features and the complexity of state-of-the-art vocoders. In the analysis part, the continuous F0 (contF0) [19] is firstly calculated that can track rapid changes without any voiced-unvoiced decision. In order to obtain the boundaries of the glottal period in the voiced regions, a glottal closure instant (GCI) algorithm [11] is employed, and then using a principle component analysis (PCA) approach to build a residual speech signal. MVF [12] is a second vocoded parameter used to separate the frequency bands into two components: periodic (low-frequency) and aperiodic (high-frequency). Finally, Mel-Generalized Cepstral analysis (24–order MGC) [56] was used to extract speech spectral information (with $gamma = -1/3$, $alpha = 0.42$, *and* $frameshift = 5\ ms$).

Drugman and Dutoit [11] have proved that the PCA based residual gives subjectively higher quality than pulse-noise excitation. Consequently, PCA residuals overlap-added pitch synchronously based on the contF0 in the synthesis part of the continuous vocoder. Next, the voiced excitation is lowpass filtered at the frequency provided by the MVF contour, whereas the white noise is selected at high frequencies. The overlap-add method is then used to synthesize the same speech signal using the Mel generalized-log spectrum approximation (MGLSA) filter [22].

2.2 Noise modeling

It was argued in [7], that the noise component of the excitation source is not accurately modeled in parametric vocoders (e.g., STRAIGHT). Also, some other vocoders lack the voiced component at higher frequencies [6]. Recently, we developed temporal envelopes (Amplitude, Hilbert, Triangular, and True) to model the noise excitation component [2]. These envelopes are neither perfect nor without their constraints, but they have been previously proven to be useful for their intended use. Therefore, we should continue our investigation of further approaches to propose a more reliable one for the continuous vocoder, as shown in Fig. 1.

2.2.1 True envelope

The True Envelope (TE) approach is a form of cepstral smoothing log amplitude spectra [45, 58]. It starts with estimating the cepstrum of the residual frame and then updating it recursively by the maximum between the spectrum of the frame signal and the current cepstral representation. In this study, the cepstrum $c(n)$ can be computed as the inverse Fourier transform of the log magnitude spectrum $S(k)$ of a speech signal frame $v(n)$

$$c(n) = \sum_{k=0}^{N-1} S(k).e^{j\left(\frac{2\pi}{N}\right)kn} \tag{1}$$

$$S(k) = log|V(k)| \tag{2}$$

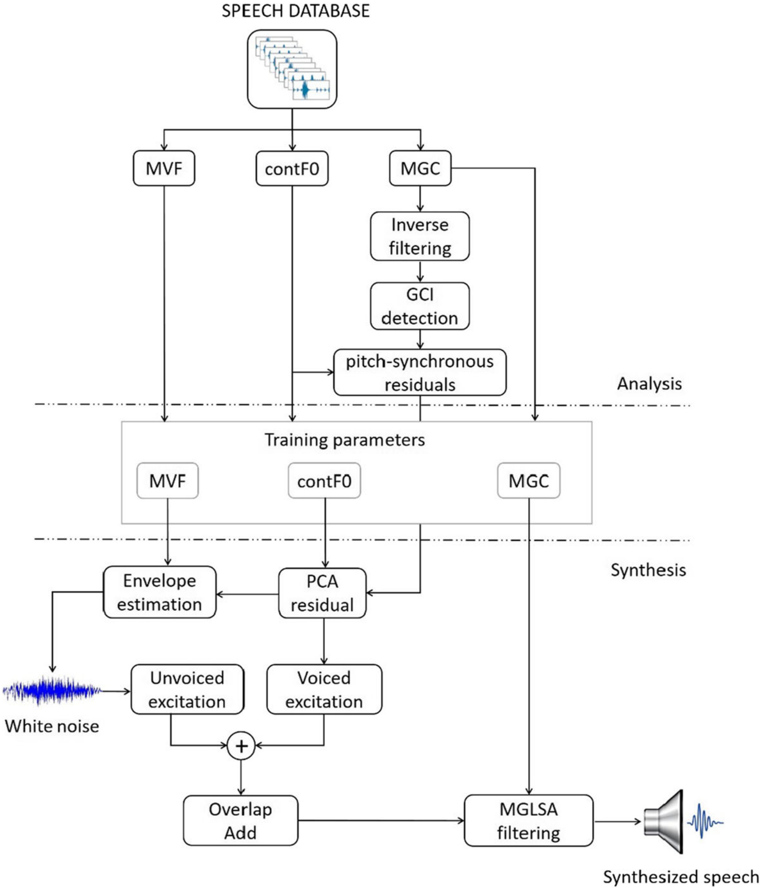


Fig. 1 Workflow of the proposed method

where $V(k)$ is N-point discrete Fourier transform of a $v(n)$ and can be estimated as

$$V(k) = \sum_{n=0}^{N-1} v(n).e^{-j\left(\frac{2\pi}{N}\right)nk} \tag{3}$$

Afterwards, the algorithm iteratively updates $M(k)$ with the maximum of the $S(k)$ and the Fourier transform of the cepstrum $C_i(k)$, that is the cepstral representation of the spectral envelope at iteration i.

$$C(k) = \sum_{n=0}^{N-1} c(n).e^{-j\left(\frac{2\pi}{N}\right)nk} \tag{4}$$

$$M_i(k) = \max(S_{i-1}(k), C_{i-1}(k)) \tag{5}$$

The TE with a weighting factor w_f would improve the temporal envelope (in practice, w_f= 10). As shown in Fig. 2, TE envelope $T(n)$ is obtained here by

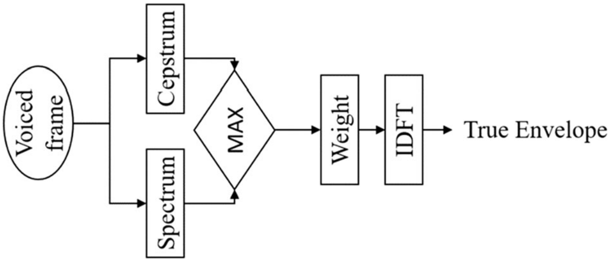


Fig. 2 True envelope estimation process

$$T(n) = \sum_{k=0}^{N-1} w_f.M(k).e^{j\left(\frac{2\pi}{N}\right)kn} \tag{6}$$

Although the performance of the TE was confirmed previously, TE creates oscillations each time the change in $S(k)$ is fast. This can be shown in Fig. 3b.

2.2.2 Discrete all-pole

Discrete all-pole envelope (DAP) is one of the most useful methods used in parametric modelling, which provide an accurate estimation of time envelope. It was proposed in [14] when only a discrete set of spectral peak points is given and used Itakura-Saito (IS) distortion measure that evaluates a matching error E_{IS} around peaks of spectral densities by iteratively minimizing

$$E_{IS} = \frac{1}{N}\sum_{k=1}^{N}\frac{P(f_k)}{\widehat{P}(f_k)} - ln\frac{P(f_k)}{\widehat{P}(f_k)} - 1 \tag{7}$$

where $P(f_k)$ is the power spectrum of the voiced signal defined at N frequency points f_m for $f = 1, 2, \ldots, N$, $\widehat{P}(f_k)$ is the power spectrum of all pole envelope. Hence, DAP envelope can be estimated at iteration m by

$$a_{k+1} = a_k(1-\delta) + \delta R^{-1}h_k \tag{8}$$

where a_k is predictor coefficients k,δ is a scalar determining the convergence speed, R is the autocorrelation of the discrete signal spectrum, and h is the time-reversed impulse response of the discrete frequency sampled all-pole model A

$$h(-n) = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{A(f_k)e^{j2\pi f_k n}} \tag{9}$$

By setting the δ to 0.5 in this paper, we can get a fast convergence rate and an error decrease every iteration. Figure 3c shows the effects of applying DAP envelope on the residual signal.

2.2.3 Frequency domain linear prediction

Another time envelope can be applied to the speech signal is the FDLP (frequency domain linear prediction), which is reliable for time-adaptive behaviour. The idea behind this technique is to use frequency-domain dual of time-domain to extract the envelope by using linear

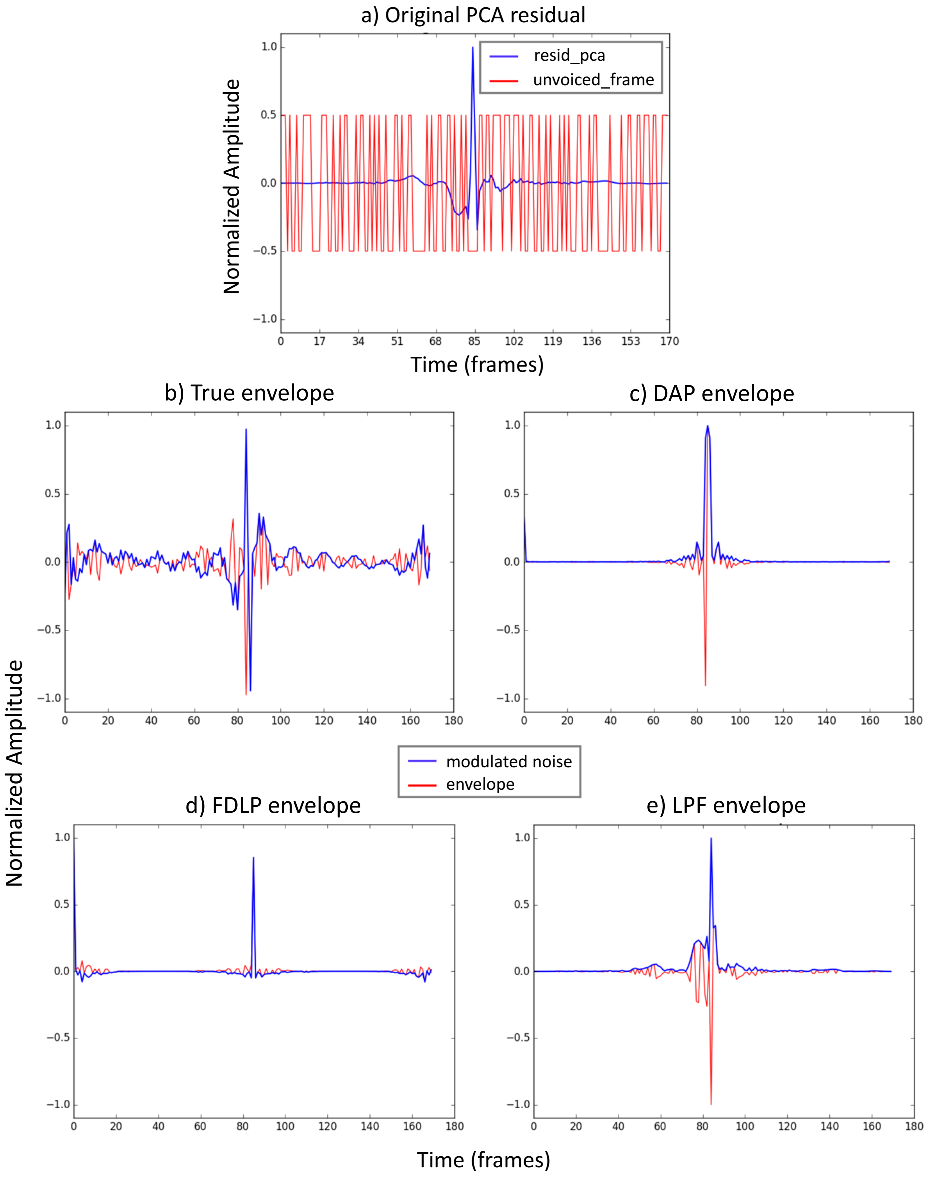


Fig. 3 Example of the temporal envelopes performance. "unvoiced_frame" is the excitation signal consisting of white noise, whereas "resid_pca" is the first eigenvector resulting from the PCA compression on the voiced excitation frame

prediction. It was formerly proved in [20] and developed in [45]. In our implementation, the complex analytic signal $c[n]$ of a discrete-time voiced frame $x[n]$ is firstly calculated through the concept of Hilbert transform $H[\cdot]$

$$c[n] = x[n] + jH[x[n]] \tag{10}$$

Next, its squared magnitude is transformed by a discrete cosine transform (DCT) to provide frequency-domain representation that is real-valued. Then, a set of Hanning overlapping windows are applied to the DCT components. After applying linear prediction, the FDLP envelope can be obtained finally by taking the inverse Fourier transform to the spectral autocorrelation coefficients a_k

$$A(n) = \frac{G}{\left|\sum_{k=0}^{P} a_k e^{-j2\pi kn}\right|^2} \tag{11}$$

where p is the FDLP model order, and G denotes the gain of the model. In our experiment, we use a model with $a_0 = 1$ to predict the current sample and setting $G = 1$ to provide better speech synthesis performance. FDLP envelope performance can be seen in Fig. 3d.

2.2.4 Low pass filtering

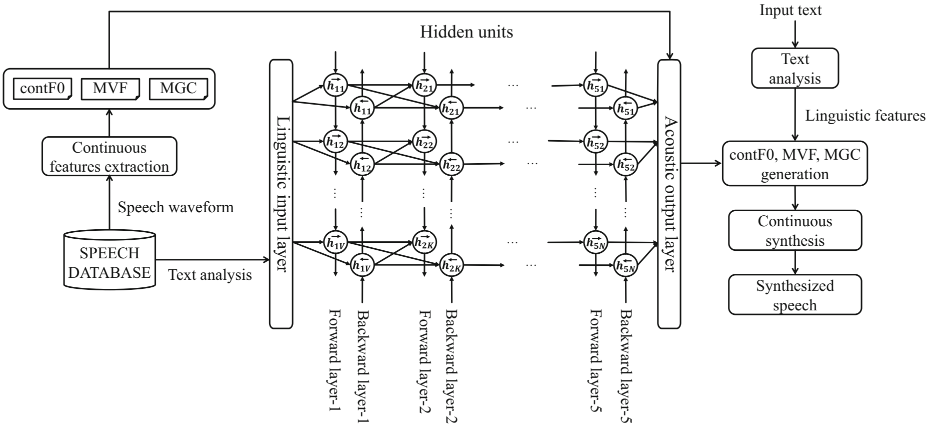
A further effective and most straightforward technique of calculating such an envelope is based on a notion of low pass filtering (LPF). The LPF envelope can be easily constructed in this paper by squaring, smoothing, and then taking the square root to the local energy of the voiced signal x[n]

$$A(n) = \sqrt{LPF\left|x[n]^2\right|} \tag{12}$$

Since low cutoff frequencies probably create an envelope with ripples, we obtain much better results through a high cutoff frequency (i.e., 4–8 kHz). Consequently, this technique brings out important information in the time that almost matches the peaks of the residual speech signal (see Fig. 3e).

2.3 Acoustic modelling within TTS

The core parts of the continuous vocoder when employed in neural network appear in Fig. 4. Textual and phonetic transcriptions are transformed to a sequence of linguistic features as input, and deep learning is used to predict acoustic features for reconstructing speech.



contF0
MVF
MGC
Continuous features extraction
Speech waveform
SPEECH DATABASE
Text analysis
Linguistic input layer
Hidden units
$\overrightarrow{h}_{11}$ $\overleftarrow{h}_{11}$ $\overrightarrow{h}_{21}$ $\overleftarrow{h}_{21}$ $\overrightarrow{h}_{51}$ $\overleftarrow{h}_{51}$
$\overrightarrow{h}_{12}$ $\overleftarrow{h}_{12}$ $\overrightarrow{h}_{22}$ $\overleftarrow{h}_{22}$ $\overrightarrow{h}_{52}$ $\overleftarrow{h}_{52}$
$\overrightarrow{h}_{1V}$ $\overleftarrow{h}_{1V}$ $\overrightarrow{h}_{2K}$ $\overleftarrow{h}_{2K}$ $\overrightarrow{h}_{5N}$ $\overleftarrow{h}_{5N}$
Forward layer-1
Backward layer-1
Forward layer-2
Backward layer-2
Forward layer-5
Backward layer-5
Acoustic output layer
Input text
Text analysis
Linguistic features
contF0, MVF, MGC generation
Continuous synthesis
Synthesized speech

Fig. 4 A schematic diagram of the proposed system for text-to-speech

2.3.1 Deep feed-forward neural network

Our baseline system [6] was successfully used with HMM-based TTS. However, HMMs often generate over-smoothing, and muffled synthesized speech. Recently, neural approaches have achieved significant improvements to replace the decision tree applied in HMM-based speech [63]. Deep neural networks (DNNs) have also shown their capability to model the high dimension of the resulting acoustic parameters [10] and to perform multi-task learning [59]. From these points, we propose a training scheme for multilayered perceptron, which tries to use the modified version of the continuous vocoder in DNN-TTS for further improving its quality. The DNN-TTS used in this work is a feed-forward multilayered architecture with six layers of hidden units, each consisting of 1024 units.

2.3.2 Recurrent neural network

Recurrent neural networks (RNNs) are another acoustic model architecture which can process sequence-to-sequence learning models. RNNs have the ability to map and store feature sequences of acoustic events, which is essential for RNN-TTS systems to enhance prediction outputs adequately. Although the proposed vocoder based DNN-TTS outperformed the baseline based HMM (see Subsection 4.3.1), Zen and Senior [61] comprehensively listed several drawbacks of the DNNs based speech synthesis, for example, lack of knowledge to predict variances and ignoring the sequential nature of speech. To cope with these issues, we study the use of sequence-to-sequence modeling with RNN based TTS. Four neural network topologies (long short-term memory (LSTM), bidirectional LSTM (Bi-LSTM), gated recurrent network (GRU), and a new hybrid RNN unit) were successfully examined and applied in TTS using the proposed vocoder.

- **Long short-term memory**

As it was introduced in [21] and applied for speech synthesis [64], LSTM is a class of RNN architecture that can hold long term memories with three gated units (input, forget, and output) to use speech data in the recurrent computations.

- **Bidirectional LSTM**

In a unidirectional RNN (Uni-RNN), only contextual information from past time instances is received, while a bidirectional RNN (Bi-RNN) can read past and future contexts by handling data in mutual directions [50]. Bi-RNN can achieve this by splitting hidden layers into forward and backward states sequence. Combining Bi-RNN with LSTM yields a bidirectional-LSTM (Bi-LSTM) that can access to long-range context [17].

- **Gated recurrent unit**

A simplified version of the LSTM is called the gated recurrent unit (GRU) architecture, which recently reached better performance than LSTM [31]. GRU has only update and reset gates to modulate the flow of data without having separate memory cells. Therefore, the total size of GRU parameters is less than that of LSTM, which lets to converge faster and prevent overfitting.

- **Hybrid model**

The advantage of RNNs is that they are capable of using the previous context. In particular, the RNN based Bi-LSTM acoustic model has been confirmed to give a state-of-the-art performance on speech synthesis tasks [1, 17]. There are two significant drawbacks to use fully Bi-LSTM hidden layers. Firstly, the speed of training becomes very slow due to iterative multiplications over time, that leads to network paralysis problems. The second problem is that the training process can be tricky and sometimes expensive task due to gradient vanishing and exploding [4].

In an attempt to overcome these limitations, we propose a modification to the fully Bi-LSTM layers by using Bi-LSTM for lower layers and conventional Uni-RNN for upper layers to reduce complexity and to make the training easier while all the contextual information from past and future have been already saved in the memory. Consequently, reducing memory requirements and the potential of being suitable for real-time applications are the main advantages of using this topology.

2.4 Speech conversion model

In [9, 30], voice conversion models based on neural networks give better performance than the GMM. Here, feed-forward DNN is applied to model the transformation of the parameters from a source to a target speaker with parallel corpora as displayed in Fig. 5. It includes three main parts: parameterization, alignment-training, and conversion-synthesis. The analysis function of the developed vocoder is used to extract the three continuous parameters from the source and target voices. A training phase based on FF-DNN is also used to build the conversion part.

The conversion function aims to map the training parameters of the source speaker $X = \{x_i\}_{i=1}^{I}$ to the corresponding one of the target speaker $Y = \left\{y_j\right\}_{j=1}^{J}$. Since these vectors differ in the duration, the Dynamic Time Warping (DTW) algorithm [41, 48] is used to align both feature vectors X and Y. DTW is a nonlinear mapping function used to minimize the total distance $D(X, Y)$ between frames of the source and target speakers. Next, a training phase based on FF-DNN is used to predict the time-aligned target features from the characteristic of the source speaker. Lastly, the converted $contF0$, MVF, and MGC are passed through the synthesis function of the Continuous vocoder to get the final converted speech waveform.

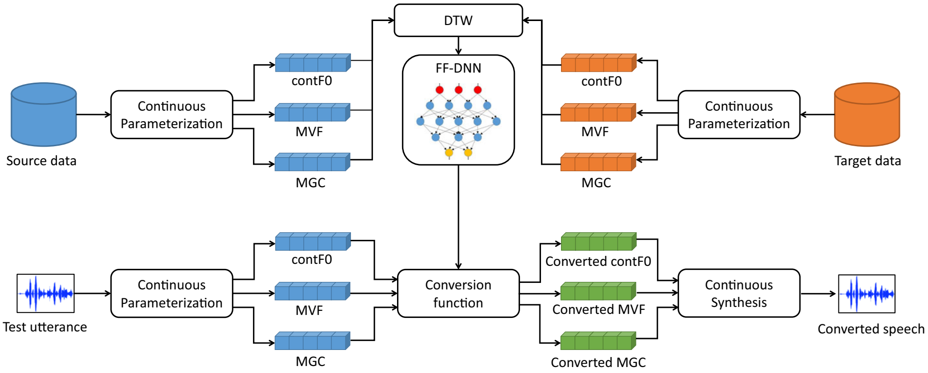


Fig. 5 Flowchart of the proposed VC algorithm

3 Experimental conditions

3.1 Datasets

The speech data is selected from a database having recordings of natural TTS synthesis. This means that three male (denoted as AWB for Scottish, BDL for American, and JMK for Canadian) English speakers with two female (denoted as SLT and CLB for US) English speakers are chosen from CMU-ARCTIC [29], each one consisting of 1132 recorded sentences. In all experiments, 90% of speech waveform files are used for training, whereas the rest (more than 100 sentences) were applied for testing with the baseline and proposed vocoders. The sampling rate of the speech database is 16 kHz.

3.2 Training topology

The general DNN/RNN model-based TTS and VC systems used in this research were applied in the open-source Merlin framework [64]. Next, constructive changes are presented to be able to adapt our continuous vocoder. A high-performance NVidia Titan X graphics processing units (GPU) was used for training the network. We trained an FF-DNN with several different sequence-to-sequence architectures, as follows:

- **FF-DNN:** 6 feed-forward hidden layers; Each layer has 1024 hyperbolic tangent units.
- **LSTM:** 3 feed-forward hidden lower layers, followed by a single LSTM hidden top layer. Each layer has 1024 hyperbolic tangent units.
- **Bi-LSTM:** It has the same architecture as LSTM, but it tends to replace only the top hidden layer with a Bi-LSTM layer.
- **GRU:** It has the same architecture as LSTM, but it tends to replace only the top hidden layer with a GRU layer.
- Hybrid: 2 Bi-LSTM hidden lower layers followed by another 2 standard RNN top layers, each of which has 1024 units.

In VC experiments, intra-gender and cross-gender pairs have been conducted. Thus, the total number of combinations is 12 pairs.

4 Experimental evaluations

To reach our aims and to prove the validity of the proposed models, objective and perceptual tests were performed.

4.1 Objective metrics for TTS

4.1.1 Phase distortion deviation

It was shown in [7] that the glottal source information could be extracted using the phase distortion. As we modeled the noise excitation component with temporal envelope, phase distortion deviation (PDD) has been computed for the vocoded samples of the proposed and baseline models. Initially, PDD was determined by using Fisher's standard deviation [18].

Nevertheless, [7] shows two more issues in variance and source shape in the voiced speech segments. Thus, by taking away from these limitations, PDD and its mean (PDM) can be estimated in this experiment at 5 *ms* frameshift by

$$PDD = \sigma_i(f) = \sqrt{-2log\left|\frac{1}{N}\sum_{n\in C} e^{j(PD_n(f)-\mu_n(f))}\right|} \tag{13}$$

$$PDM = \mu_i(f) = \angle\left(\frac{1}{N}\sum_{n\in C} e^{jPD_n(f)}\right) \tag{14}$$

where $C = \left\{i-\frac{N-1}{2}, \ldots, i+\frac{N-1}{2}\right\}$, N is the frame number, PD is the phase difference between two consecutive frequencies, and we indicate the phase by $\angle$.

In this experiment, the phase distortion values below the MVF contour have been attenuated (zeroed out) in order to measure only the noisiness in the high-frequency regions. For this, we plotted in Fig. 6 the PDD of one original (natural) sample with four vocoded (synthesized) variants. It can be clearly seen that the 'FDLP' vocoding sample has noticeably different noisy components compared to the original sentence (the colour differences between 1 s and 2 s). In contrast, the proposed approach with 'LPF' envelope has PDD values closer to that of the original speech and superior to the baseline envelope (e.g., the colour differences between 0.5–1.5 s).

In addition, we calculated the distribution of the PDD measure and Mann-Whitney-Wilcoxon ranksum tests for numerous sentences. In doing so, the phase distortion means for the male and female speakers categorized by the five variants is presented in Fig. 7. In view of that, the PDD values of the 'baseline' system are close to natural speech. The system with 'LPF' envelope is matched to the natural speech. The 'DAP' and 'FDLP' envelopes result in different PDD values, but in general, they are further away from the natural speech for the female speaker. We must now conclude, therefore, that the True and LPF envelopes are suitable for noise modeling in our continuous vocoder.

Besides our contribution in [2], this further supports the claim that by applying one of the above temporal envelope estimation methods and parameterizing the noise component in the continuous vocoder, the system produces better results than using the vocoder without envelopes.

4.1.2 RMS – Log spectral distance

In our recent studies [2, 6], 24–order mel-generalized cepstral coefficients [56] were used to represent the spectral model in the continuous vocoder. Besides, some advanced spectral approaches possibly can improve the quality of synthetic speech. Therefore, Cheaptrick algorithm [34] using 60-order MGC representation will be used to achieve the desired speech spectral quality in the TTS and VC systems.

Several performance indices have been proposed for evaluating spectral algorithms. Since we are dealing with speech synthesis based TTS and one major task is the refinement of the spectral envelopes in our continuous vocoder, we will concentrate on distance measures. Spectral distortion is among the most popular ones and plays an essential role in the assessment of the speech quality, which is designed to compute the distance between two

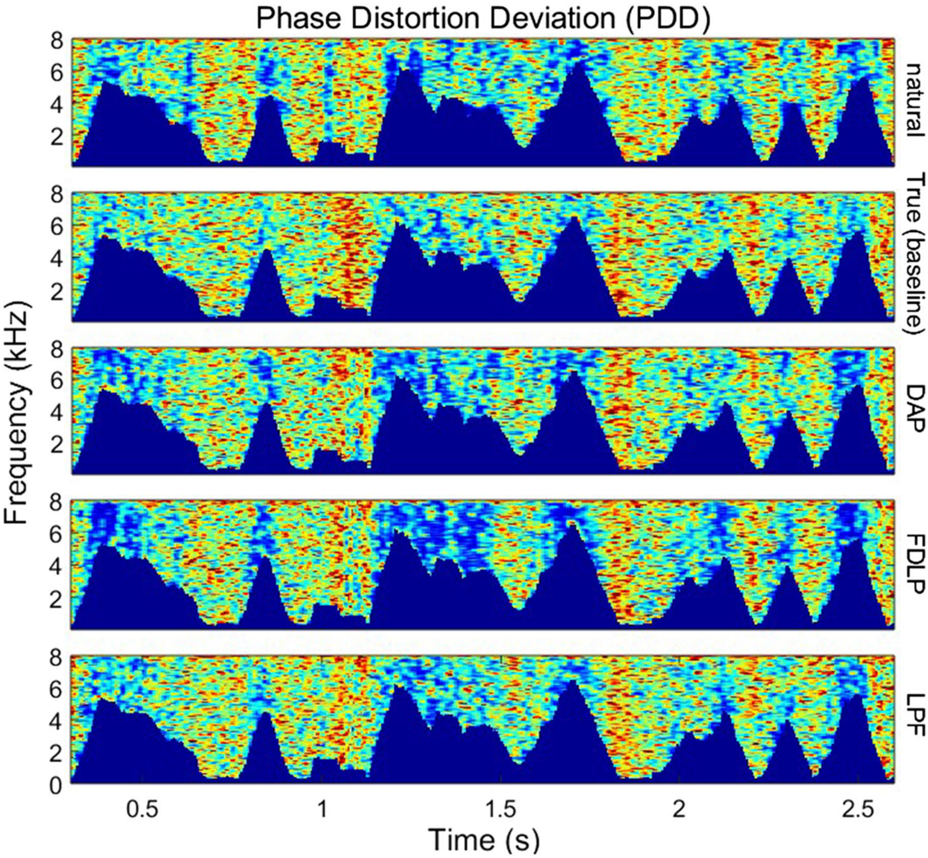


Fig. 6 Phase Distortion Deviation of vocoded speech samples (sentence: "Jerry was so secure in his nook that he did not roll away." from speaker AWB). The warmer the color, the bigger the PDD value and the noisier the corresponding time-frequency region

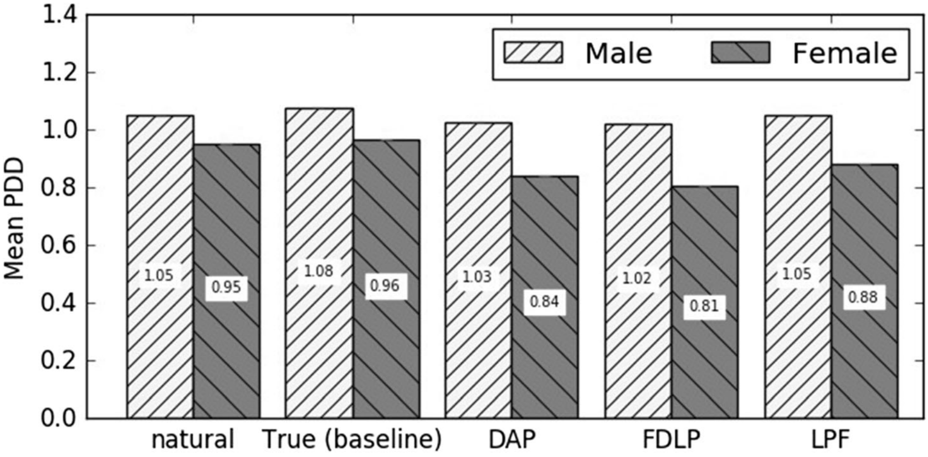


Fig. 7 Values of the Mean PDD by sentence type

power spectra. Therefore, the root mean square (RMS) log spectral distance (LSD) metric is proposed here to carry out the evaluation in this study by

$$LSD_{RMS} = \sqrt{\frac{1}{N}\sum_{k=1}^{N} mean\left[logP(f_k)-log\widehat{P}(f_k)\right]^2} \tag{15}$$

where $P(f)$ is the spectral power magnitudes of the original speech, while $\widehat{P}(f_k)$ is the spectral power magnitudes of the synthesized speech. The optimal value of $\mathrm{LSD_{RMS}}$ is zero, which indicates a matching frequency content. The values expressed in Table 1 refer to the average $\mathrm{LSD_{RMS}}$ that was calculated for 20 sentences selected randomly from two categories of SLT and AWB speakers. The analysis of these results confirms that the $\mathrm{LSD_{RMS}}$ for the CheapTrick algorithm is better than the standard one used in the baseline vocoder. Moreover, Fig. 8 showed three spectrograms of frequency versus time. In the middle spectrogram, the $\mathrm{LSD_{RMS}}$ of the signal is equal to 1.6, while the bottom spectrogram has a lower $\mathrm{LSD_{RMS}}$ equal to 0.89 that is closer to the top speech spectrogram (natural speech). Consequently, our proposed vocoder introduces a smaller distortion to the sound quality and approaches a correct spectral criterion.

4.1.3 Comparison of the WORLD and continuous vocoders

The WORLD vocoder was chosen for comparison with our optimized vocoder for the reason that it also used a CheapTrick spectral algorithm. Initially, the WORLD vocoder was proposed in [36]. Similarly to the continuous vocoder, the WORLD vocoder is based on source-filter separation, i.e. models the spectral envelope and excitation separately (with F0 and aperiodicity).

It can be noted from Table 2 that the proposed vocoder takes an only 1–dimensional parameter for forming two excitations, while the WORLD system is using a 5-dimensional band aperiodicity.

Moreover, the F0 modeling capability and the voiced/unvoiced (V/UV) transitions were examined between continuous and WORLD vocoders. Even though the WORLD vocoder can give better quality when applied in TTS, it can make V/UV decision errors. For example, setting voiced that should be unvoiced or setting unvoiced that should be voiced, and occasionally contains errors at transition boundaries (V/UV or UV/V). For that reason, the V/UV error was 5.35% for the WORLD synthesizer in case of the female speaker. This is not the case with the continuous vocoder, which is using a continuous F0 algorithm, and the continuous Maximum Voiced Frequency parameter models the voicing feature. Therefore, V/UV errors do not happen in our system. In view of that, F0 contour of a synthesized speech sample using the DIO (Distributed Inline-filter Operation) algorithm [35] and the contF0 are shown in Fig. 9.

Table 1 Average log spectral distance for the spectral estimation

Spectral algorithm	$\mathrm{LSD_{RMS}}$ (dB)	
	SLT	AWB
Standard MGC	1.47	0.94
CheapTrick MGC	**0.91**	**0.89**

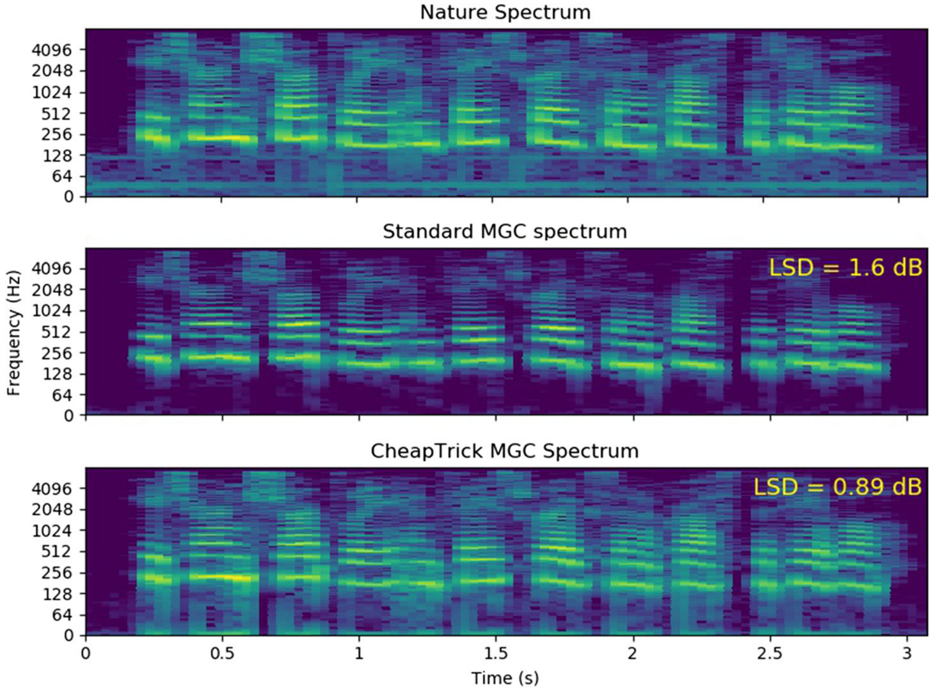


Fig. 8 Comparison of speech spectrograms. The sentence is "He turned sharply, and faced Gregson across the table.", from speaker SLT

4.1.4 Comparison of the deep learning architectures using empirical measures

To obtain a clear picture of how these four recurrent networks perform against the DNN, five metrics have been used to assess their performance:

1. **MCD (dB):** 60-dimensional mel-cepstral distortion coefficients.
2. **MVF (Hz):** Root mean squared error to measure the performance of the maximum voiced frequency.
3. **F0 (Hz):** Root mean squared error to measure the performance of the fundamental frequency.
4. **Overall validation error:** A validation loss between valid and train sets from last epoch (iteration).
5. **CORR:** The correlation measures the degree to which reference and generated contF0 data are close to each other (linearly related).

Table 2 Parameters of applied vocoders

Vocoder	Parameter per frame	Excitation
Continuous	F0: 1 + MVF: 1 + MGC: 60	Mixed
WORLD	F0: 1 + Band aperiodicity: 5 + MGC: 60	Mixed

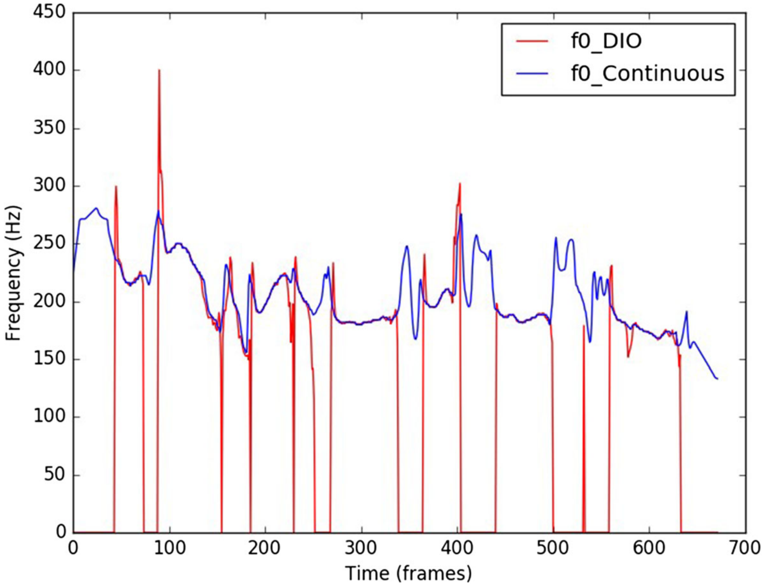


Fig. 9 F0 trajectories of a synthesized speech signal using the DIO algorithm (red), and continuous algorithm (blue) for continuous and WORLD vocoders respectively

For all experimental metrics, the computation is done frame-by-frame, and a smaller value shows better performance except for the CORR measure where +1 is better. Overall validation error throughout the training decreases with epochs, which indicates a convergence. The test results for the proposed recurrent models are listed in Table 3. Compared to the FF-DNN, the Bi-LSTM reduces all four experimental measures, and obtain similar performance for the male and female speakers. Although the Hybrid system is not better than Bi-LSTM, it slightly drops the validation error in case of AWB speaker from 1.632 in Bi-LSTM to 1.627. Besides, the Hybrid system does not outperform the baseline model. This indicates that increasing the number of recurrent units in the hidden layers are not helpful. We also see that using GRU system has no positive effect on the objective metrics. In summary, these empirical outcomes prove that using Bi-LSTM methods to train the parameters of the developed vocoder enhances the synthesis performance and superior to the feed-forward DNN and other recurrent topologies.

Table 3 Objective measures for all training systems based on vocoded samples using proposed Continuous vocoder for SLT and AWB speakers

Systems	MCD (dB)		MVF (Hz)		F0 (Hz)		CORR		Validation error	
	SLT	AWB	SLT	AWB	SLT	AWB	SLT	AWB	SLT	AWB
FF-DNN	4.923	4.592	0.044	0.046	17.569	22.792	0.727	0.803	1.543	1.652
LSTM	4.825	4.589	0.046	0.047	17.377	23.226	0.732	0.793	1.526	1.638
GRU	4.879	4.649	0.046	0.047	17.458	23.337	0.731	0.791	1.529	1.643
Bi-LSTM	**4.717**	**4.503**	**0.042**	**0.044**	**17.109**	**22.191**	**0.746**	**0.809**	**1.517**	1.632
Hybrid	5.064	4.516	0.046	0.044	18.232	22.522	0.704	0.805	1.547	**1.627**

Bold font indicates the best values

4.2 Objective metrics for VC

In this work, our developed model has been compared with three high-quality systems recently suggested for VC: WORLD [36], MagPhase [16], and Sprocket [28]. We ran our model over 48 experiments, and two important metrics are considered for testing and validating the performance of the suggested system. These are Frequency-weighted segmental signal-to-noise ratio (fwSNRseg) [32] and Log-Likelihood Ratio (LLR) [44]. fwSNRseg defined as

$$\mathrm{fwSNR_{seg}} = \frac{1}{\mathrm{N}} \sum_{\mathrm{j=1}}^{\mathrm{N}} \left(\frac{\sum_{\mathrm{i=1}}^{\mathrm{K}} \mathrm{W_{i,j}} \cdot \log \frac{\mathrm{X_{i,j}^2}}{\mathrm{X_{i,j}^2 - Y_{i.j}^2}}}{\sum_{\mathrm{i=1}}^{\mathrm{K}} \mathrm{W_{i,j}}} \right) \tag{16}$$

where X, Y are magnitude spectra of the target and converted speech signals respectively, and W is a weight vector. Whereas LLR determine the distance between two signals, which takes the form

$$\mathrm{LLR} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i=1}}^{\mathrm{N}} \log \left(\frac{\mathrm{a_{y,i}^T \; R_{x,i} \; a_{y,i}}}{\mathrm{a_{x,i}^T \; R_{x,i} \; a_{x,i}}} \right) \tag{17}$$

where $\mathrm{a_x}$, $\mathrm{a_y}$, are respectively the linear prediction coefficients of the target and converted signals, and the autocorrelation matrix of the target speech signal is $\mathrm{R_x}$.

A calculation is performed frame-by-frame, and the scores were averaged over 20 synthesized samples for each conversion pair as shown in Table 4. It can be shown firstly that the proposed model is significantly preferred over other methods in female-to-male speech and female-to-female conversions (i.e. SLT-to-BDL or CLB-to-SLT). Moreover, fwSNRseg measure tended to have the highest scores with the developed vocoder in male-to-female conversions (i.e. BDL-to-SLT, BDL-to-CLB, and JMK-to-CLB). While in male-to-male voice conversion (i.e. JMK-to-BDL), the proposed system yields the second-highest score.

Further, we compared the continuous parameters between the source, target, and converted speech sentences as displayed in Fig. 10. It may be noticed that both converted and target

Table 4 Average performance scores

Model	WORLD		MagPhase		Sprocket		Proposed	
	fwSNRseg	LLR	fwSNRseg	LLR	fwSNRseg	LLR	fwSNRseg	LLR
BDL → JMK	2.19	1.57	**3.21**	**1.37**	2.20	1.48	2.47	1.50
BDL → SLT	1.12	1.72	1.25	1.69	1.04	**1.49**	**2.33**	1.57
BDL → CLB	0.79	1.83	1.65	1.72	0.37	**1.69**	**1.66**	1.74
JMK → BDL	1.31	1.76	**2.49**	**1.56**	1.73	1.63	2.15	1.57
JMK → SLT	0.55	1.74	**1.93**	**1.56**	0.11	1.64	1.54	1.65
JMK → CLB	1.45	1.74	1.75	1.66	0.69	**1.60**	**1.81**	1.67
SLT → BDL	1.65	1.71	1.60	1.70	1.80	1.51	**2.95**	**1.49**
SLT → JMK	2.16	1.61	**2.71**	1.42	0.713	1.56	2.59	**1.39**
SLT → CLB	1.51	1.75	**2.89**	1.59	2.32	1.56	2.51	**1.50**
CLB → BDL	0.97	1.81	1.60	1.70	0.95	1.72	**1.92**	**1.60**
CLB → JMK	2.50	1.49	2.74	1.40	0.98	1.46	**3.00**	**1.30**
CLB → SLT	0.98	1.70	**2.17**	1.53	1.96	1.54	2.12	**1.47**

Bold font indicates the best values

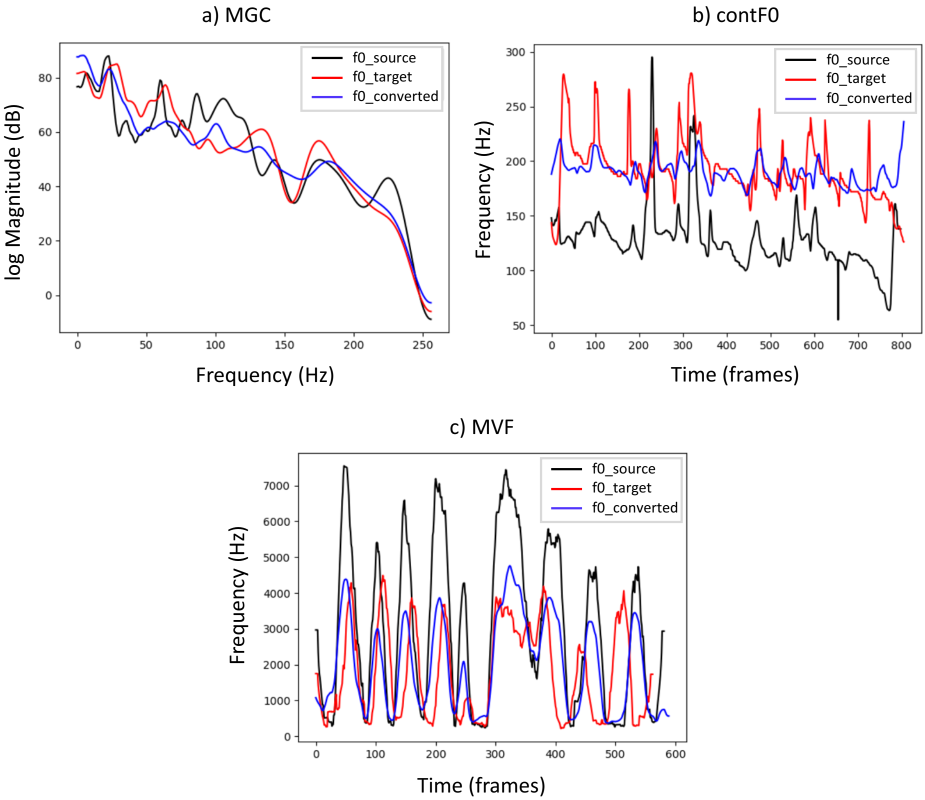


Fig. 10 Example of the natural source, natural target, and converted continuous contours using the proposed approach

spectral envelopes are mostly similar in Fig. 10a than the source one. Similarly, in Fig. 10b, the converted contF0 contour almost give the same shape of the target contour, which can produce high-quality F0 predictions. It can also be seen in Fig. 10c that the converted parameter of the MVF is more similar to the target than to the source. Overall, these experiments demonstrate that the proposed model with continuous vocoder is competitive for the voice conversion task.

4.3 Subjective evaluations

Here, we perform three listening tests based on TTS and VC systems, the samples of which can be found online.[1]

4.3.1 Listening test #1: DNN-TTS

First, in order to evaluate the differences in DNN-TTS synthesized samples using the above vocoders, we ran an online MUSHRA (MUlti-Stimulus test with Hidden Reference and Anchor) perceptual test [23]. We evaluated natural utterances with the synthesized ones from the baseline, WORLD, proposed, HMM-TTS [6], and an anchor system (pulse-noise excitation

[1] http://smartlab.tmit.bme.hu/vocoder2019

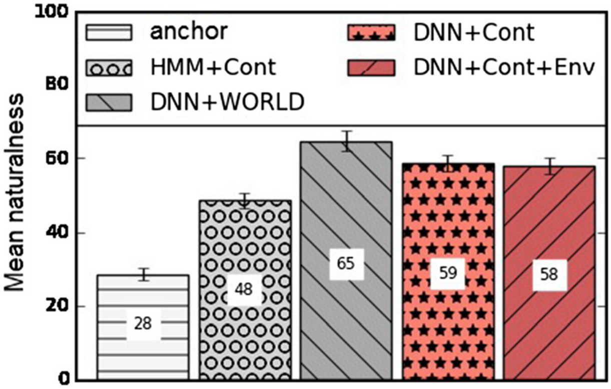


Fig. 11 Scores of the MUSHRA listening test #1. Error bars display the bootstrapped 95% confidence intervals

vocoder). Fifteen sentences from a female speaker were chosen randomly, which means that ninety wave files were put in the MUSHRA test (6 models × 15 sentences). Listeners had to judge the naturalness of each utterance compared to the natural sentence, from 0 (extremely unnatural) to 100 (extremely natural). The sentences appeared in a randomized order.

Nine applicants were requested to complete the listening test (seven males, two females). The assessment took twenty minutes to fill, on average. Figure 11 shows the results of this test based on DNN-TTS. Accordingly, the DNN-TTS with the continuous vocoder significantly outperformed the baseline method based on HMM-TTS, whereas its naturalness is slightly worse than the WORLD vocoder.

4.3.2 Listening test #2: RNN-TTS

In the second listening assessment, we evaluated natural sentences against the synthesized sentences from the baseline (FF-DNN), proposed (Bi-LSTM, Hybrid), and an HMM-TTS (anchor system) using a simple pulse-noise excitation vocoder. As we noticed insignificant differences in quality between LSTM and GRU samples and to make the test easier to the subjects, we only included samples from Hybrid and Bi-LSTM systems. We assessed ten sentences randomly selected from speaker AWB, and ten sentences from speaker SLT.

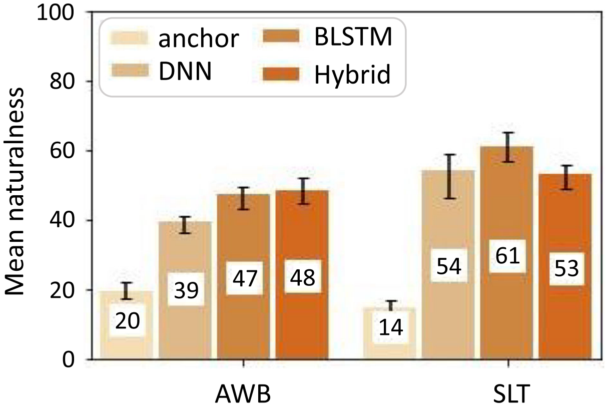


Fig. 12 Scores of the MUSHRA listening test #2. Error bars display the bootstrapped 95% confidence intervals

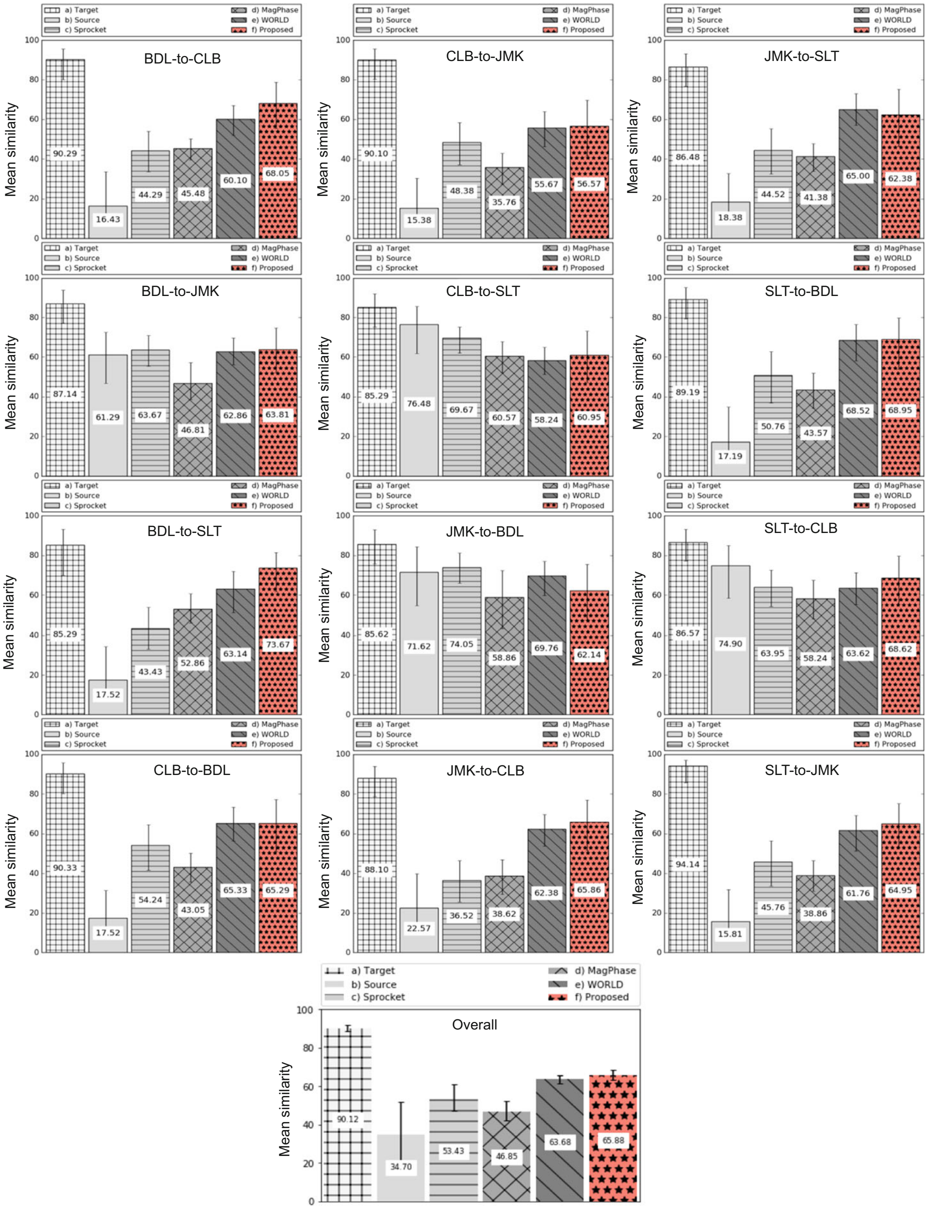


Fig. 13 MUSHRA scores for the similarity question. Higher value means better overall quality. Errorbars show the bootstrapped 95% confidence intervals

Another 13 participants with an engineering background (6 males, 7 females) were invited to conduct the online perceptual test which took 23 min to fill. Figure 12 shows the results of the MUSHRA scores. For speaker AWB, both recurrent networks outperformed the FF-DNN system, and the Bi-LSTM and Hybrid networks are not significantly different from each other (Mann-Whitney-Wilcoxon ranksum test, $p < 0.05$). For speaker SLT, we found that the Bi-

LSTM method gives the highest naturalness scores, and this difference is statistically significant between the Bi-LSTM and Hybrid systems.

4.3.3 Listening test #3: VC

In this section, we design a MUSHRA-like listening test to assess the similarity of the converted voice to a natural target one. The number of participants in this test is 20 (11 males and 9 females), in which they evaluated 72 utterances (6 types × 12 sentences selected randomly). The time length of the MUSHRA test is roughly 10 min, and the samples can be found online.[2] The similarity ratings of the MUSHRA test are displayed in Fig. 13.

The results show that the overall accuracy of the proposed system in the similarity test is better than others developed earlier. This means the suggested VC model executed by the continuous vocoder has effectively converted the source voice to the preferred voice on both gender cases. However, VC based on WORLD vocoder has a higher score only in the JMK-to-SLT case; whereas Sprocket also has higher scores only in the JMK-to-BDL and CLB-to-SLT cases. Finally, to sum up the results obtained from this test, the proposed system is recommended as the best conversion technique, while the WORLD system is a second good option.

5 Conclusion

This article suggested firstly a new method for modelling unvoiced sounds in a continuous vocoder by estimating a proper time envelope. Three different envelopes (DAP, FDLP, and LPF) were created, tested, and compared to the baseline vocoder. From the objective measurements using Phase Distortion Deviation (PDD), we found that the True and LPF envelopes give better results when applied in the continuous vocoder (they are almost matching the original utterances in terms of PDD) than other envelopes. Secondly, we build a deep learning model based TTS to increase the quality of synthesized speech and train all continuous parameters used by the vocoder. The motivation behind this experiment arose from our examination that the state-of-the-art WORLD vocoder frequently has V/UV errors and boundary errors because of its fundamental frequency algorithm. Moreover, the DNN-TTS based proposed vocoder was rated as better than an earlier HMM-TTS approach in a MUSHRA listening test. A further goal reported in this study was to focus on the recurrent neural networks based on the continuous vocoder. LSTM, Bi-LSTM, GRU, and Hybrid models are applied to train our parameters, and experimental results proved that our suggested recurrent Bi-LSTM model could increase the naturalness of the synthesized speech significantly. Finally, a new statistical voice conversion approach was proposed based on deep learning. The strengths and weaknesses of the proposed method were examined using a variety of measurement techniques. From the objective and subjective tests, the proposed system based on VC was mostly superior to the other approaches.

Consequently, the benefit of the continuous vocoder is straightforward, flexible, having only two 1-dimensional parameters, and its synthesis algorithm is computationally feasible for real-time implementation. Besides, the continuous vocoder does not have a voiced/unvoiced decision, which means that the alignment error between voiced and unvoiced segments will always be avoided in VC compared to the conventional techniques. For future work, the authors plan to investigate the non-parallel many-to-many voice conversion based on convolution neural networks.

[2] http://smartlab.tmit.bme.hu/vocoder2019

Acknowledgements The research was partly supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825619 (AI4EU), and by the National Research Development and Innovation Office of Hungary (FK 124584 and PD 127915). The Titan X GPU used was donated by NVIDIA Corporation. We would like to thank the subjects for participating in the listening test.

Funding Open access funding provided by Budapest University of Technology and Economics.

References

1. Al-Radhi MS, Csapó TG, Németh G (2017) Deep Recurrent Neural Networks in Speech Synthesis Using a Continuous Vocoder. In: proceedings of the speech and computer (SPECOM), Hatfield, UK, Springer, 2017, p 282–291. https://doi.org/10.1007/978-3-319-66429-3_27
2. Al-Radhi MS, Csapó TG, Németh G (2017) Time-domain envelope modulating the noise component of excitation in a continuous residual-based vocoder for statistical parametric speech synthesis. In: proceedings of the Interspeech, Stockholm, p 434-438. https://doi.org/10.21437/Interspeech.2017-678
3. Athineos M, ELLIS D (2003) Frequency-domain linear prediction for temporal features. Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, USA, In, pp 261–266. https://doi.org/10.1109/ASRU.2003.1318451
4. Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw 5(2):157–166. https://doi.org/10.1109/72.279181
5. Cabral JP, Berndsen JC (2013) Towards a better representation of the envelope modulation of aspiration noise. In: proceedings of the advances in nonlinear speech processing (NOLISP), Berlin Heidelberg, p 67-74. https://doi.org/10.1007/978-3-642-38847-7_9
6. Csapó TG, Németh G, Cernak M (2015) Residual-based excitation with continuous F0 modeling in HMM-based speech synthesis. In: Proceedings of the international conference on statistical language and speech processing. Budapest, Hungary, pp 27–38. https://doi.org/10.1007/978-3-319-25789-1_4
7. Degottex G, Erro D (2014) A uniform phase representation for the harmonic model in speech synthesis applications. EURASIP Journal of Audio, Speech, Music Processing 38(1):1–16. https://doi.org/10.1186/s13636-014-0038-1
8. Degottex G, Lanchantin P, Gales M (2016) A pulse model in log-domain for a uniform synthesizer. In: proceedings of the international speech communication association speech synthesis workshop (ISCA SSW9), California, p 230–236
9. Desai S, Raghavendra EV, Yegnanarayana B, Black AW, Prahallad K (2009) Voice conversion using artificial neural networks. In: proceedings of the international conference on acoustics, speech, and signal processing (ICASSP), Taiwan, p 3893-3896. https://doi.org/10.1109/ICASSP.2009.4960478
10. Ding C, Xie L, Zhu P (2015) Head motion synthesis from speech using deep neural networks. Multimed Tools Appl 74:9871–9888. https://doi.org/10.1007/s11042-014-2156-2
11. Drugman T, Dutoit T (2012) The deterministic plus stochastic model of the residual signal and its applications. IEEE Trans Audio Speech Lang Process 20(3):968–981. https://doi.org/10.1109/TASL.2011.2169787
12. Drugman T, Stylianou Y (2014) Maximum voiced frequency estimation: exploiting amplitude and phase spectra. IEEE Signal Processing Letters 21(10):1230–1234. https://doi.org/10.1109/LSP.2014.2332186
13. Drullman R (1995) Temporal envelope and fine structure cues for speech intelligibility. J Acoust Soc Am 97(1):585–591. https://doi.org/10.1121/1.413112
14. El-Jaroudi A, Makhoul J (1991) Discrete all-pole modeling. IEEE Trans Signal Process 39(2):411–423. https://doi.org/10.1109/78.80824

15. Espic F, Botinhao CV, King S (2017) Direct Modelling of magnitude and phase spectra for statistical parametric speech synthesis. In: proceedings of the Interspeech, Stockholm, p 1383-1387. https://doi.org/10.21437/Interspeech.2017-1647
16. Espic F, Valentini-Botinhao C, King S (2017) Direct Modelling of magnitude and phase spectra for statistical parametric speech synthesis. In: proceedings of the Interspeech, Stockholm, p 1383-1387. https://doi.org/10.21437/Interspeech.2017-1647
17. Fan Y, Qian Y, Xie F (1964-1968) Soong FK (2014) TTS synthesis with bidirectional LSTM based recurrent neural networks. In, Proceedings of the Interspeech, Singapore, p
18. Fisher NI (1995) Statistical analysis of circular data. Cambridge University, UK. https://doi.org/10.1017/CBO9780511564345
19. Garner PN, Cernak M, Motlicek P (2013) A simple continuous pitch estimation algorithm. IEEE Signal Processing Letters 20(1):102–105. https://doi.org/10.1109/LSP.2012.2231675
20. Herre J, Johnston JD (1996) Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS). In: proceedings of the 101st audio engineering society convention, Los Angeles.
21. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
22. Imai S, Sumita K, Furuichi C (1983) Mel log Spectrum approximation (MLSA) filter for speech synthesis. Electronics and Communications in Japan (Part I: Communications) 66(2):10–18. https://doi.org/10.1002/ecja.4400660203
23. ITU-R BS.1534 (2001) Method for the subjective assessment of intermediate audio quality.
24. Kai Y, Steve Y (2011) Continuous F0 modeling for HMM based statistical parametric speech synthesis. IEEE Trans Audio Speech Lang Process 19(5):1071–1079. https://doi.org/10.1109/TASL.2010.2076805
25. Kai Y, Blaise T, Steve Y (2010) From discontinuous to continuous F0 Modelling in HMM-based speech synthesis. In: proceedings of the international speech communication association speech synthesis workshop (ISCA SSW7), Kyoto, p 94–99
26. Kaneko T, Kameoka H, Hiramatsu K, Kashino K (2017) Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks. In: proceedings of the Interspeech, Stockholm, p 1283-1287. https://doi.org/10.21437/Interspeech.2017-970
27. Kawahara H, Katsuse IM, Cheveigne AD (1999) Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction. Speech Comm 27(3–4):187–207. https://doi.org/10.1016/S0167-6393(98)00085-5
28. Kobayashi K, Toda T (2018) Sprocket: open-source voice conversion software. In: proceedings of the odyssey: the speaker and language recognition, p 203-210. https://doi.org/10.21437/Odyssey.2018-29
29. Kominek J, Black AW (2003) CMU ARCTIC databases for speech synthesis. Language Technologies Institute [Online] Available: http://festvox.org/cmu_arctic/
30. Kotani G, Saito D, Minematsu N (2017) Voice conversion based on deep neural networks for time-variant linear transformations. In: proceedings of the Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC), Kuala Lumpur, p 1259-1262. https://doi.org/10.1109/APSIPA.2017.8282216
31. Li X, Yuan A, Lu X (2018) Multi-modal gated recurrent units for image description. Multimed Tools Appl 77:29847–29869. https://doi.org/10.1007/s11042-018-5856-1
32. Ma J, Hu Y, Loizou PC (2009) Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. Acoustical Society of America 125(5):3387–3405. https://doi.org/10.1121/1.3097493
33. Maia R, Zen H, Knill K, Gales M (1833–1836) Buchholz S (2011) multipulse sequences for residual signal modeling. In, Proceedings of the Interspeech, p
34. Morise M (2015) CHEAPTRICK, a spectral envelope estimator for high-quality speech synthesis. Speech Comm 67:1–7. https://doi.org/10.1016/j.specom.2014.09.003
35. Morise M, Kawahara H, Nishiura T (2010) Rapid f0 estimation for high SNR speech based on fundamental component extraction. IEICE transactions information systems J93-D(2):109–117
36. Morise M, Yokomori F, Ozawa K (2016) WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. IEICE Trans Inf Syst E99-D(7):1877–1884. https://doi.org/10.1587/transinf.2015EDP7457
37. Najafabadi M, Villanustre F, Khoshgoftaar T, Seliya N, Wald R, Muharemagic E (2015) Deep learning applications and challenges in big data analytics. Journal of Big Data 2(1):1–21. https://doi.org/10.1186/s40537-014-0007-7
38. Nakamura K, Toda T, Saruwatari H, Shikano K (2012) Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech. Speech Comm 54(1):134–146. https://doi.org/10.1016/j.specom.2011.07.007
39. Nakashika T, Takashima R, Takiguchi T, Ariki Y (2013) Voice conversion in high-order eigen space using deep belief nets. In: Proceedings of the Interspeech. Lyon, France, pp 369–372

40. Nakashika T, Takiguchi T, Ariki Y (2014) Voice conversion based on speaker-dependent restricted boltzmann machines. IEICE Trans Inf Syst 97(6):1403–1410
41. Ney H (1984) The use of a one-state dynamic programming algorithm for connected word recognition. IEEE Trans Acoust Speech Signal Process 32(2):263–271. https://doi.org/10.1109/TASSP.1984.1164320
42. Pantazis Y, Stylianou Y (2008) Improving the modeling of the noise part in the harmonic plus noise model of speech. In: proceedings of the international conference on acoustics, speech, and signal processing (ICASSP), Nevada, p 4609-4612. https://doi.org/10.1109/ICASSP.2008.4518683
43. Qiong H, Richmond K, Yamagishi J, Latorre J (2013) An experimental comparison of multiple vocoder types, in: proceedings of the international speech communication association speech synthesis workshop (ISCA SSW8), Barcelona, p 155–160
44. Quackenbush S, Barnwell T, Clements M (1988) Objective measures of speech quality. Prentice-Hall, Englewood Cliffs, NJ
45. Robel A, Rodet X (2005) Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation. Proceedings of the International Conference on Digital Audio Effects, Spain, In, pp 30–35
46. Robel A, Villavicencio F, Rodet X (2007) On cepstral and all-pole based spectral envelope modeling with unknown model order. Pattern Recogn Lett 28(1):1343–1350. https://doi.org/10.1016/j.patrec.2006.11.021
47. Saito Y, Ijima Y, Nishida K, Takamichi S (2018) Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors. In: proceedings of the international conference on acoustics, speech, and signal processing (ICASSP), Canada, p 5274-5278. https://doi.org/10.1109/ICASSP.2018.8461384
48. Sakoe H, Chiba S (1978) Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions Acoustic Speech Signal Processing 26(1):43–49. https://doi.org/10.1109/TASSP.1978.1163055
49. Schloss A (1985) On the automatic transcription of percussive music - from acoustic signal to high-level analysis, Ph.D. thesis, Stanford University.
50. Schuster M, Paliwal K (1997) Bidirectional recurrent neural networks. IEEE Trans Signal Process 45(11): 2673–2681. https://doi.org/10.1109/78.650093
51. Stylianou Y (2001) Applying the harmonic plus noise model in concatenative speech synthesis. IEEE Transactions Audio, Speech, Language Processing 9(1):21–29. https://doi.org/10.1109/89.890068
52. Stylianou Y, Cappe O, Moulines E (1998) Continuous probabilistic transform for voice conversion. IEEE Transactions Speech and Audio Processing 6(2):131–142. https://doi.org/10.1109/89.661472
53. Sun L, Kang S, Li K, Meng H (2015) Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. In: proceedings of the international conference on acoustics, speech, and signal processing (ICASSP), Brisbane, Australia, 2015, pp. 4869-4873. https://doi.org/10.1109/ICASSP.2015.7178896
54. Toda T, Black AW, Tokuda K (2007) Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. IEEE Trans Audio Speech Lang Process 15(8):2222–2235. https://doi.org/10.1109/TASL.2007.907344
55. Toda T, Nakagiri M, Shikano K (2012) Statistical voice conversion techniques for body-conducted unvoiced speech enhancement. IEEE Trans Audio Speech Lang Process 20(9):2505–2517. https://doi.org/10.1109/TASL.2012.2205241
56. Tokuda K, Kobayashi T, Masuko T, Imai S (1994) Mel-generalized cepstral analysis - a unified approach to speech spectral estimation. In: proceedings of the international conference on spoken language processing (ICSLP), Yokohama, p 1043–1046.
57. Tokuda K, Mausko T, Miyazaki N, Kobayashi T (2002) Multi-space probability distribution HMM. IEICE Trans Inf Syst E85-D(3):455–464
58. Villavicencio F, Robel A, Rodet X (2006) Improving LPC spectral envelope extraction of voiced speech by true-envelope estimation. In: Proceedings of the international conference on acoustics, speech, and signal processing (ICASSP). Toulouse, France, pp 869–872. https://doi.org/10.1109/ICASSP.2006.1660159
59. Wu Z, Botinhao CV, Watts O, King S (2015) Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In: Proceedings of the international conference on acoustics, speech, and signal processing (ICASSP). Brisbane, Australia, pp 4460–4464. https://doi.org/10.1109/ICASSP.2015.7178814
60. Wu Z, Chng ES, Li H (2015) Exemplar-based voice conversion using joint nonnegative matrix factorization. Multimed Tools Appl 74:9943–9958. https://doi.org/10.1007/s11042-014-2180-2
61. Zen H, Senior A (2014) Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In: proceedings of the international conference on acoustics, speech, and signal processing (ICASSP), Florence, p 3844-3848. https://doi.org/10.1109/ICASSP.2014.6854321
62. Zen H, Tokuda K, Black AW (2009) Statistical parametric speech synthesis. Speech Comm 51(11):1039–1064. https://doi.org/10.1016/j.specom.2009.04.004

63. Zen H, Senior A, Schuster M (2013) Statistical parametric speech synthesis using deep neural network. In: proceedings of the international conference on acoustics, speech, and signal processing (ICASSP), Vancouver, p 7962–7966. https://doi.org/10.1109/ICASSP.2013.6639215
64. Zhizheng W, Watts O, King S (2016) Merlin: an open source neural network speech synthesis system. In: proceedings of the international speech communication association speech synthesis workshop (ISCA SSW9), Sunnyvale, CA, USA. https://doi.org/10.21437/SSW.2016-33

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Mohammed Salah Al-Radhi was born in Basra, Iraq. He got a BSc degree in Computer Engineering at Basra University in 2007, and a MSc degree in Communication Systems at Portsmouth University, UK which was achieved with first honors in 2012 and awarded the MSc top student certificate in 2013. He started from September 2016 to complete his PhD at the Speech Technology and Smart Interactions Laboratory in the Budapest University of Technology and Economics. He is working on designing vocoders and acoustic models for statistical speech synthesis. His main interests are the signal processing, speech synthesis, machine learning, acoustic models, and voice conversion.

Tamás Gábor Csapó obtained his MSc in computer science from Budapest University of Technology and Economics (BME), Hungary in 2008. Between 2008 and 2014, he was a doctoral student at the Speech Technology and Smart Interactions Laboratory of BME, where he obtained his PhD degree. In 2007, he was awarded with 1st prize of the National Conference of Scientific Student's Associations, Hungary. He received a CIRE student grant of the Acoustical Society of America in 2010 and was a

Fulbright scholar at Indiana University, USA in 2014, where he started to deal with ultrasound tongue imaging. In 2016, he joined the MTA-ELTE Lingual Articulation Research Group, focusing on investigating the Hungarian articulation during speech production. Since 2017, he has two national research projects about ultrasound-based silent speech interfaces. His research interests include speech synthesis, speech analysis, vocoding and ultrasound-based tongue movement analysis.

Géza Németh was born in 1959. He obtained his MSc in electrical engineering, major in Telecommunications at the Faculty of Electrical Engineering of BME in 1983. Also at BME: dr. univ., 1987, PhD 1997. He is an associate professor at BME. He is the author or co-author of more than 170 scientific publications and 4 patents. His research fields include speech technology, service automation, multilingual speech and multimodal information systems, mobile user interfaces and applications. He is the Head of the Speech Technology and Smart Interactions Laboratory of BME TMIT.