



# Continuous vocoder applied in deep neural network based voice conversion

Mohammed Salah Al-Radhi<sup>1</sup>  · Tamás Gábor Csapó<sup>1,2</sup> · Géza Németh<sup>1</sup>

Received: 26 January 2019 / Revised: 10 July 2019 / Accepted: 6 September 2019 /  
Published online: 16 September 2019

© The Author(s) 2019

## Abstract

In this paper, a novel vocoder is proposed for a Statistical Voice Conversion (SVC) framework using deep neural network, where multiple features from the speech of two speakers (source and target) are converted acoustically. Traditional conversion methods focus on the prosodic feature represented by the discontinuous fundamental frequency (F0) and the spectral envelope. Studies have shown that speech analysis/synthesis solutions play an important role in the overall quality of the converted voice. Recently, we have proposed a new continuous vocoder, originally for statistical parametric speech synthesis, in which all parameters are continuous. Therefore, this work introduces a new method by using a continuous F0 (contF0) in SVC to avoid alignment errors that may happen in voiced and unvoiced segments and can degrade the converted speech. Our contribution includes the following. (1) We integrate into the SVC framework the continuous vocoder, which provides an advanced model of the excitation signal, by converting its contF0, maximum voiced frequency, and spectral features. (2) We show that the feed-forward deep neural network (FF-DNN) using our vocoder yields high quality conversion. (3) We apply a geometric approach to spectral subtraction (GA-SS) in the final stage of the proposed framework, to improve the signal-to-noise ratio of the converted speech. Our experimental results, using two male and one female speakers, have shown that the resulting converted speech with the proposed SVC technique is similar to the target speaker and gives state-of-the-art performance as measured by objective evaluation and subjective listening tests.

**Keywords** Voice conversion · Continuous vocoder · neural network · Speech synthesis

---

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

Tamás Gábor Csapó  
csapot@tmit.bme.hu

Géza Németh  
nemeth@tmit.bme.hu

Extended author information available on the last page of the article

## 1 Introduction

Statistical Voice Conversion (SVC) is a potential technique to enable a user for flexibly synthesizing several kinds of speech. While keeping the linguistic content and environmental conditions unchanged, the goal of SVC is to change and modify speaker individuality; i.e., the source speaker's voice is transformed to sound like that of the target speaker [7]. There are several applications within the concept of voice conversion, such as converting speech from impaired to normal voice [11], from normal to singing sound [47], electro-laryngeal to normal speech [44], etc.

Over the years, voice conversion frameworks have mostly focused on spectral conversion between source and target speakers [8, 21]. In the sense of the statistical parametric approaches, such as Gaussian mixture model (GMM) [59] and exemplar based on non-negative matrix factorization [1, 63], SVC showed a success in the linear transformation of the spectral information. Nonlinear transformation approaches, such as hidden Markov models (HMMs) [49], deep belief networks (DBNs) [45] and restricted Boltzmann machines (RBMs) [46], have been also shown to be effective in modeling the relationship between source-target features more accurately. The DBN and RBM were used to replace GMM to model the distribution of spectral envelopes [36]. However, the resulting speech parameters from these models tend to be over-smoothed and affect the similarity and quality of generated speech. To cope with these problems, some approaches attempt to reduce the difference between natural and the converted speech parameters by using Global variance [59], modulation spectrum [57], dynamic kernel partial least squares regression [20], or generative adversarial networks [28]. Even though these techniques achieve some improvements, the naturalness of the converted voice still deteriorates compared to the source speaker. Therefore, improving the performance of converted voice is still a challenging research question.

There seem to be four factors that degrade the quality of SVC: 1) speech parameters (i.e. vocoder features), 2) mapping function between the source and target speakers, 3) learning model, and 4) vocoder synthesis quality. To capture the quality of these factors, feed-forward deep neural networks (FF-DNNs) was proposed as an acoustic modeling solution of different research areas [19, 50, 66]. FF-DNNs have shown their ability to extract high-level, complex abstractions and data representations from large volumes of supervised and unsupervised data [43], and achieve significant improvements in various machine learning areas including the ability to model high-dimensional acoustic parameters [61], and the availability of multi-task learning [64]. In this article, we predict acoustic features using a FF-DNN, which are then passed to a vocoder to generate the converted speech waveform. Thus, both vocoder and FF-DNN models can be used to improve the converted acoustic parameters.

A vocoder (which is also called speech analysis/synthesis system) is another important component of various speech synthesis applications such as Text-To-Speech (TTS) synthesis [16], voice conversion [31], or singing synthesizers [29]. Although there are several different types of vocoders, they follow the same main strategy. The analysis stage is used to convert speech waveform into a set of parameters which represent separately the vocal-folds excitation signal and the vocal-tract filter transfer function to filter the excitation signal, whereas in the synthesis stage, the entire parameter set is used to reconstruct the original speech signal. Hu et al. [23] present an experimental comparison of a wide range of important vocoder types which have been previously invented. Despite the fact that most of these vocoders have been successful in synthesizing speech, the sound quality of the synthesized voices is still perceptibly degraded compared to that of the natural sound. The reason for this is either the inaccurate

estimation of the vocoder parameters that lead to losing some important excitation / spectral details, or typically those vocoders are computationally intensive. However, various studies in voice conversion are still considering some of these vocoders [23], such as STRAIGHT [5, 53, 59], mixed excitation [34], Harmonic plus Noise Model [35], glottal source modeling [6], or even with more complex end-to-end acoustic models like adaptive WAVENET [54], or Tacotron [62]. Consequently, simple and uniform vocoders, which would handle all speech sounds and voice qualities (e.g. creaky voice) in a unified way, are still missing in SVC. Therefore, it is still worth to develop advanced vocoders for achieving high-quality converted speech.

In our recent work in statistical parametric speech synthesis, we have proposed a novel continuous vocoder using continuous fundamental frequency (contF0) in combination with Maximum Voiced Frequency (MVF), which was shown to improve the performance under a FF-DNN compared to the hidden Markov model based TTS [2]. The advantage of a continuous vocoder in this scenario is that vocoder parameters are simpler to model than in traditional vocoders with discontinuous F0. However, in SVC, the effectiveness of the continuous vocoder has not been confirmed yet. Thus, we are developing a solution in this article to achieve higher sound quality and conversion accuracy, while the SVC remains computationally efficient.

Unlike the methods referenced above, the proposed structure implicates two major technical developments. First, we build a voice conversion framework that consists of a FF-DNN and a continuous vocoder to automatically estimate the mapping relationship between the parameters of the source and target speakers. Second, we apply a geometric approach to spectral subtraction (GA-SS) to improve the signal-to-noise ratio of the converted speech and enhance anti-noise property of our vocoder. For the first time, we study the interaction between continuous parameters and FF-DNN based voice conversion. We expect that the new voice conversion model gives high-quality synthesized speech compared to the source voice.

This paper is organized as follows: In Section 2, we propose the novel idea of continuous vocoder based voice conversion. In Section 3, experimental conditions and error metrics are addressed. We report the objective and subjective evaluation results in Section 4. Section 5 gives the conclusion and discussion.

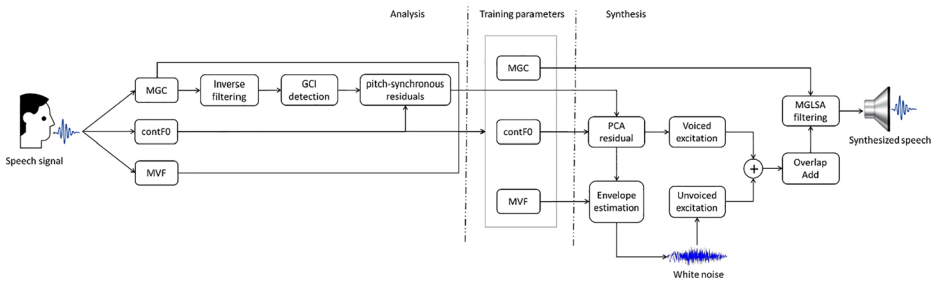
## 2 Proposed conversion methodology

### 2.1 Speaker-adaptive continuous vocoder

To construct our proposed method based SVC, we adopt a continuous F0 (contF0) estimator with maximum voiced frequency (MVF) as they are the base features for our continuous vocoder, so that they can appropriately synthesize high quality speech. The continuous vocoder was designed to overcome shortcomings of discontinuity in the speech parameters and the computational complexity of modern vocoders. Our proposed vocoder is presented in Fig. 1, and its algorithms are briefly explained below.

#### 2.1.1 contF0: F0 estimation algorithm

In recent years, there has been a rising trend of assuming that continuous F0 observations are present similarly in unvoiced regions and there have been various modelling schemes along



**Fig. 1** Workflow of the continuous vocoder

these lines. It was found in [27] that a continuous F0 creates more expressive F0 contours with HMM. Zhang et al. [67] introduce a new approach to improve modeling piece-wise continuous F0 trajectory with voicing strength and voiced/unvoiced decision for HMM-based TTS.

The contF0 estimator used in this vocoder is an approach proposed by Garner et al. [18] that is able to track fast changes. According to Fig. 2, the algorithm starts simply with splitting the speech signal into overlapping frames. The result of windowing each frame is then used to calculate the autocorrelation function. Identifying a peak between two frequencies and calculating the variance are the essential steps of the Kalman smoother to give a final sequence of continuous pitch estimates with no voiced/unvoiced decision.

Besides, during the analysis phase, the Glottal Closure Instant (GCI) algorithm [15] is used to find the glottal period boundaries of individual cycles in the voiced parts of the inverse filtered residual signal. From these pitch cycles, a Principal Component Analysis (PCA) residual is built which will be used in the synthesis phase as shown in Fig. 1 to yield better speech quality than those of the excitation pulses. In our previous study, Tóth and Csapó [60] have shown that contF0 contour can be approximated better with HMM and deep neural network (DNN) than traditional discontinuous F0. An example of contF0 estimation on a female speech sample is shown in Fig. 3 compared with the DIO algorithm [41] as one of the most successful discontinuous F0 s.

### 2.1.2 MVF: Maximum voiced frequency algorithm

During the production of voiced sounds, MVF is used as the spectral boundary separating low-frequency periodic and high-frequency aperiodic components. MVF has been used in various speech models [12, 17, 56], that yield sufficiently better quality in synthesized speech. Our vocoder follows the algorithm proposed by [13] which has the potential to discriminate harmonicity, exploits both amplitude and phase spectra, and use the maximum likelihood criterion as a strategy to derive the MVF estimate. The performance of this algorithm has been previously assessed by comparing it with two state-of-the-art methods, namely the Peak-to-Valley (P2V) used in [56] and the Sinusoidal Likeness Measure (SLM) [17]. Based on Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC), the algorithm proposed by [13] objectively outperforms both P2V and SLM methods. Moreover, a substantial improvement was also observed over the state-of-the-art techniques in a subjective listening test using male, female, and child speech.

The method consists of the following steps. First, 4 period-long Hanning window is applied to exhibit a good peak structure. Then, the frequencies of the spectral peaks are detected using a standard peak picking function. Amplitude spectrum, phase coherence, and harmonic-to-

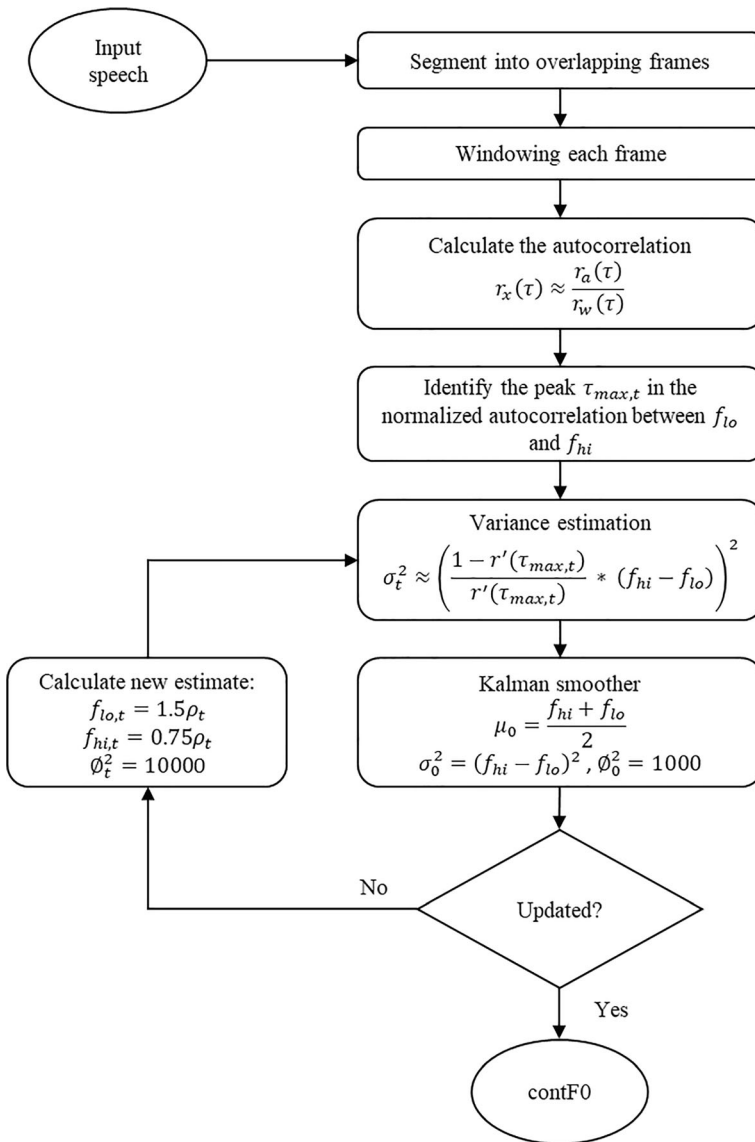
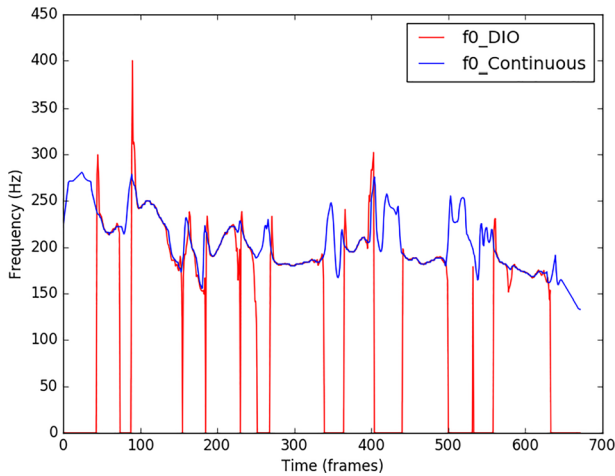


Fig. 2 Flowchart of the continuous pitch estimation algorithm

noise ratio are extracted in the third step for each harmonic candidate which convey some relevant statistics to predict the strategy decision by using the maximum likelihood criterion. Time smoothing step is finally applied to the obtained MVF trajectory in order to remove unwanted spurious values. An example of spectrogram of the natural waveform with the MVF contour is shown in Fig. 4. Here, the duration of this sentence is about 3 s, and was sampled at 16 kHz with a 16-bit quantization level. It is windowed by Hanning window function in duration of 25 ms, mutually shifted by 5 ms. The thresholds for the pitch tracking are set from 80 to 300 Hz. Thus, the MVF parameter models the voicing information: for unvoiced sounds, the MVF is low (around 1 kHz), for voiced sounds, the MVF is high (above 4 kHz).

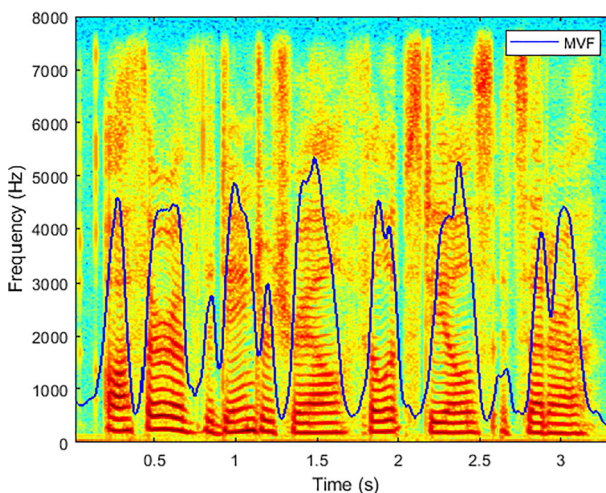


**Fig. 3** Example of F0 estimated by contF0 algorithm (blue) and DIO algorithm (red). Sentence: “Author of the danger trail, Philip Steels, etc.”, from a female speaker

### 2.1.3 MGC: Mel-generalized cepstral algorithm

In our recent studies [3, 9], a simple spectral model represented by 24-order MGC was used [58]. Although several vocoders based on this simple algorithm have been developed, they are not able to synthesis natural sound. The main problem is that it is affected by time-varying components and it is difficult to remove them. Therefore, more advanced spectral estimation methods might increase the quality of synthesized speech.

In [39], an accurate and temporally stable spectral envelope estimation called CheapTrick was proposed. CheapTrick consists of three steps: F0-adaptive Hanning window, smoothing of the power spectrum, and spectral recovery in the quefrequency domain. In a modified version of the continuous vocoder, Cheaptrick algorithm using the 60-order MGC representation with



**Fig. 4** Example of spectrogram of the natural waveform and MVF contour (blue). Sentence: “Author of the danger trail, Philip Steels, etc.”, from a female speaker

$\alpha = 0.58$  ( $F_s = 16$  kHz) will be used to achieve high-quality speech spectral estimation. A comparison of spectral envelope between standard MGC and the CheapTrick is shown in Fig. 5. Accordingly, it is clear now to see how a continuous vocoder will behave after adaptation to a more accurate spectral envelope technique than the MGC previous system.

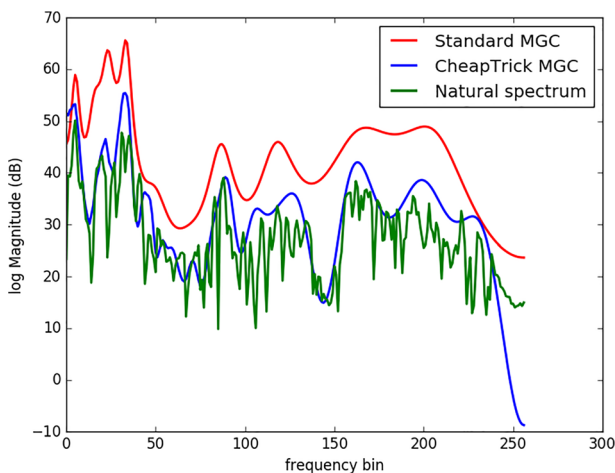
### 2.1.4 Synthesis algorithm

It was shown in [14] that the PCA based residual yields better speech quality than pulse-noise excitation. Therefore, voiced excitation in the continuous vocoder is composed of PCA residuals overlap-added pitch synchronously based on the contF0. Then, the voiced excitation is lowpass filtered frame by frame at the frequency given by the MVF parameter. In the frequencies higher than the actual value of MVF, white noise is used. The voiced and the unvoiced excitation are added together. Finally, a Mel generalized-log spectrum approximation (MGLSA) filter [24] is used to synthesize speech from the excitation and the MGC parameter stream.

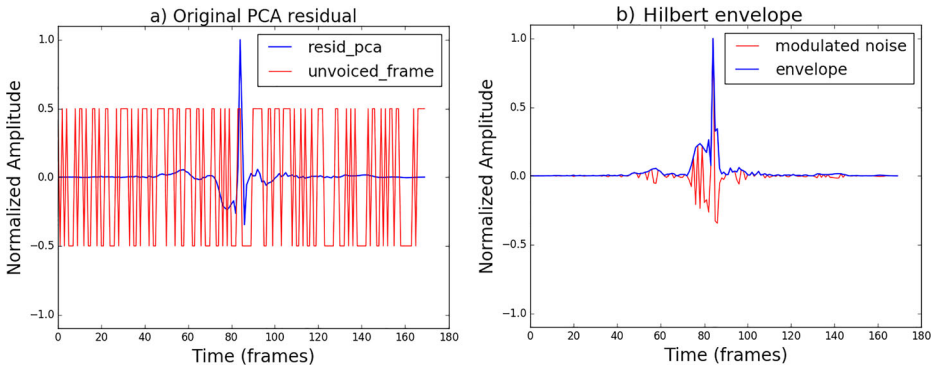
In a recent study, we applied various time envelopes to shape the high-frequency component (above MVF) of the excitation by estimating the envelope of the PCA residual; that is helpful in achieving accurate approximations compared to natural speech. In this work, we also added a time domain envelope to the voiced and unvoiced excitation to make it more similar to the residual of natural speech. This technique, using a Hilbert envelope, brings out some hidden information more efficiently and fits a curve that approximately matches the peaks of the residual frame as shown in Fig. 6. As a consequence, the analysis and synthesis steps for the latest version of the continuous vocoder are shown in Fig. 7.

## 2.2 Training a model based on FF-DNN

In [10, 38], the neural network based SVC reaches higher performance on the conversion than the GMM alternative. In this work, a FF-DNN is used to model the transformation between source and target speech features as shown in Fig. 8. It consists of 6 feed-forward hidden



**Fig. 5** Example of the signal spectrum of a voiced segment (green) with the spectral shape (spectral envelope) estimates obtained with standard MGC (red) and CheapTrick (blue)



**Fig. 6** Illustration of the performance of the time envelope. “unvoiced\_frame” is the excitation signal consisting of white noise, whereas “resid\_pca” is the result of applying PCA on the voiced excitation frames

layers, each consisting of 1024 units and performs a non-linear function of the previous layer’s representation, and a linear activation function at the output layer. These layers perform the following transformation

$$y_i(x) = f \left( \sum_{j=0}^{M_i} W_{ij}x_j + b_i \right) \tag{1}$$

where  $M_i$  is the number of units in layer  $i$ ,  $x = (x_1, \dots, x_n)$  is the input feature vector,  $y = (y_1, \dots, y_k)$  is the output vector,  $W$  is the connection weight matrix between two layers,  $b$  is the bias vector, and  $f(\cdot)$  denotes an activation function which is defined as:

$$f(x) = \begin{cases} \frac{e^{2x}-1}{e^{2x}+1}, & \text{in the hidden layer} \\ x, & \text{in the output layer} \end{cases} \tag{2}$$

FF-DNN aims to minimize the mean squared error function between the target output  $y$  and the prediction output  $\hat{y}$

$$E = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \tag{3}$$

Hence, input features are propagated forward through the FF-DNN with these estimated parameters to produce the corresponding output parameters.

1. contF0 = Continuous(X, Fs)
2. MGC = CheapTrick(X, Fs, F0)
3. MVF = MaxVoiceFreq(X, F0)
4. Time envelope = Hilbert(PCA-residual frame)
5. Y = Synthesis(contF0, MVF, MGC, Time envelope)

**Fig. 7** Steps of the continuous vocoder.  $X$  represents the input waveform,  $F_s$  represents the sampling frequency, and  $Y$  represents the synthesized speech



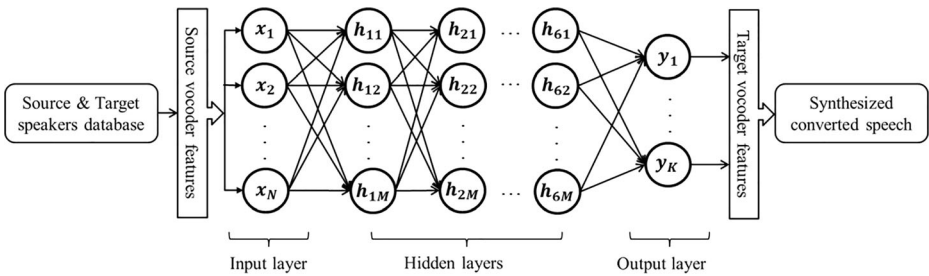


Fig. 8 A general schematic diagram of the proposed method based on FF-DNN

### 2.3 Voice conversion model

The framework of the proposed SVC system is shown in Fig. 9. It consists of feature processing, training and conversion-synthesis steps. MVF, contF0, and MGC parameters are extracted from the source and target voices using the analysis function of the continuous vocoder. A training process based on a FF-DNN is applied to construct the conversion phase.

The purpose of the conversion function is to map the training features of the source speaker  $X = \{x_i\}_{i=1}^I$  to the corresponding training features of the target speaker  $Y = \{y_j\}_{j=1}^J$ . Here,  $X$  and  $Y$  vector sequences are time-aligned frame by frame by the Dynamic Time Warping (DTW) algorithm [48, 52] since both vectors differ in the durations and have different-length recordings. DTW is a technique for deriving a nonlinear mapping between two vectors to minimize the overall distance  $D(X, Y)$  between the source and target speakers.

Then, the time-aligned acoustic feature sequences of both speakers are trained and used for the conversion function in order to predict the target features from the features of the source speaker. Finally, the converted *contF0*, *MVF*, and *MGC* are synthesized to get the converted speech waveform by the synthesis function of the continuous vocoder.

### 2.4 Reducing unwanted frequencies

The goal of this section is to remove or reduce the level of unwanted high-frequency components from the converted features, that may be generated during training or conversion phase. Therefore, we apply the GA-SS approach proposed by [65] in order to improve the performance of the converted speech signal. This approach consistently outperforms other conventional spectral subtractions particularly at low SNRs. Besides, GA-SS more suitable for our work because of its simplicity and low computational cost. Here, GA-SS can be applied in each frame signal  $f(n)$  by letting  $y(n) = f(n) + e(n)$  be the sampled speech signal with the estimation error  $e(n)$ , assuming that the first 3 frames are noise/silence. Taking the short-time Fourier transform of  $y(n)$

$$Y(w_k) = F(w_k) + E(w_k) \tag{4}$$

where  $w_k = 2\pi k/N$ ,  $k = 0, 1, 2, \dots, N - 1$ , and  $N$  is the frame length in samples. Then, we can rewrite Eq. (4) in polar form as

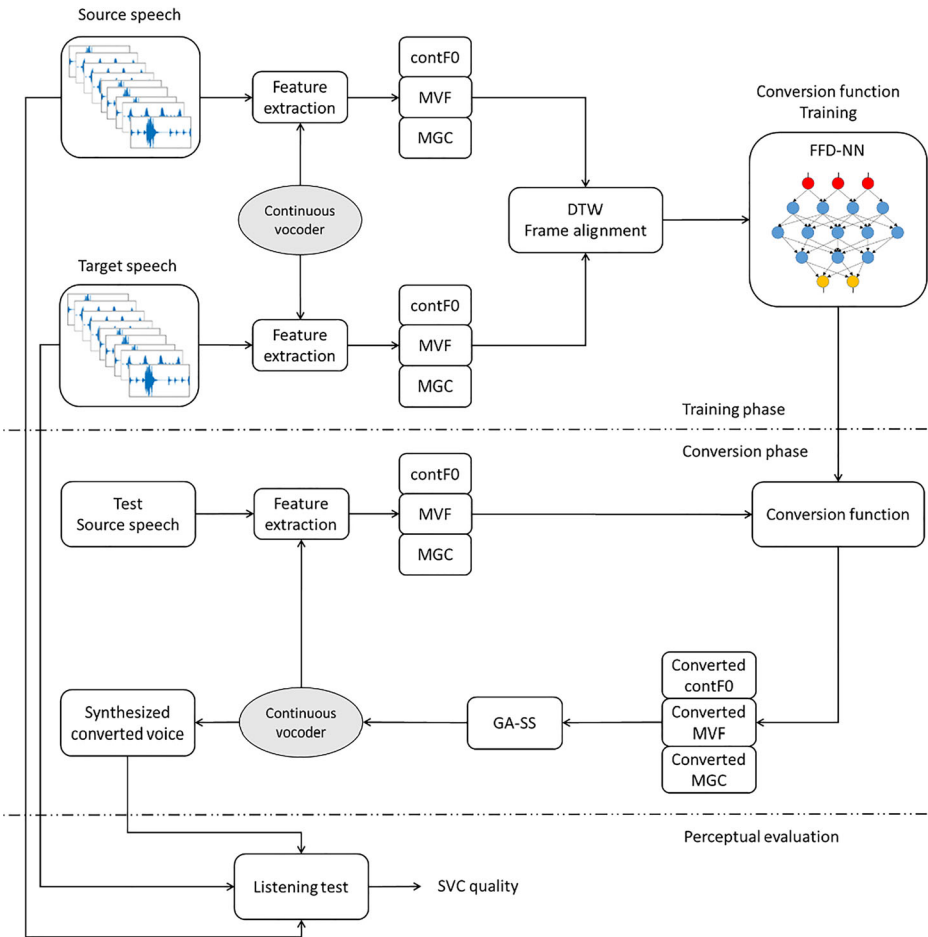


Fig. 9 Flowchart of the proposed SVC algorithm

$$A_Y e^{j\theta_Y} = A_F e^{j\theta_F} + A_E e^{j\theta_E} \tag{5}$$

where  $A$  and  $\theta$  are the magnitude and phase of the frame spectra respectively. Taking into account the trigonometric principles in Equation 5, the gain function  $H_G$  can be derived as always real and positive [65].

$$H_G = \frac{A_F}{A_Y} = \sqrt{\frac{1 - \cos^2(\theta_Y - \theta_E)}{1 - \cos^2(\theta_F - \theta_E)}} \quad , (H_G \geq 0) \tag{6}$$

Obtain the enhanced magnitude spectrum of the signal by

$$\hat{A}_F = H_G * A_Y \tag{7}$$

Using the inverse discrete Fourier transform of  $\hat{A}_F \cdot e^{j\theta_Y}$ , the enhanced frame signal  $\hat{f}(n)$  can be obtained.

To clarify the effects of this approach, white Gaussian noise is added to the natural and synthetic speech waveforms. The amount of noise is specified by signal-to-noise ratio (SNR) in the range of  $-20$  to  $10$  dB. The root mean square (RMS) error was calculated over 20 sentences selected randomly from each speaker. The smaller the value of RMS, the better performance. The overall RMS error values obtained as a function of the SNR between clean speech (natural or synthesized) sample and the noisy one (the same speech sample, with noise added) is shown in Fig. 10. The results suggest that the RMS for the synthesized signal with GA-SS approach is smallest and close to the natural signal than without GA-SS. Nevertheless, the differences were very small. But adding this approach as an extra step to our proposed model does help to some extent in improving the overall sound quality, especially in noisy conditions.

### 3 Experimental conditions

In order to evaluate the performance of the suggested voice conversion framework, a database containing a few hours of speech from several speakers was required for giving indicative results. Datasets are described in more detail in the first part of this section, while training settings and error metrics are defined afterword.

#### 3.1 Datasets

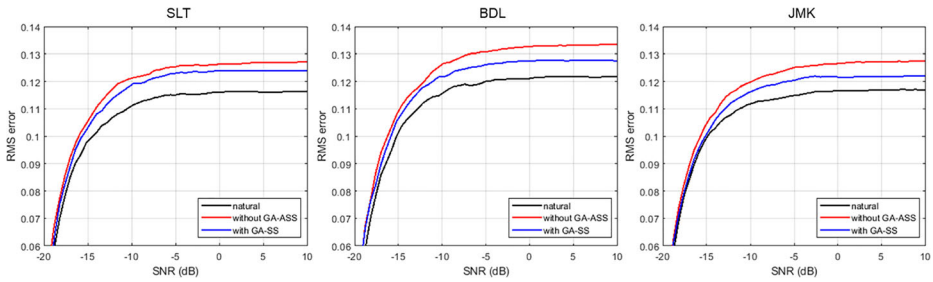
Three English speakers were chosen from the CMU-ARCTIC<sup>1</sup> database [32], denoted BDL (American English, male), JMK (Canadian English, male), and SLT (American English, female), each one consisting of 1132 sentences. The speech waveform of this database was recorded at a 16 kHz sampling rate with 16-bit linear quantization. 90% of these sentences were used in the training experiment, while the rest were used for testing and evaluating the SVC. We used SLT, BDL, and JMK for source and target speakers as well. With the aim of seeing the statistical behavior of our proposed model, we considered only cross-gender (“male-to-female” and “female-to-male”) conversions in this experiment as we do not see much of a difference between the converted and target speech in the intra-gender (“male-to-male” and “female-to-female”) conversions. Hence, four SVC experiments are carried out for evaluation:

- SLT to BDL
- BDL to SLT
- SLT to JMK
- JMK to SLT

#### 3.2 FF-DNN settings

A hyperbolic tangent activation function was applied. The outputs lie in the range  $(-1$  to  $1)$  and this function can yield lower error rates and faster convergence than a logistic sigmoid function. For the first 15 epochs, a fixed learning rate of 0.002 was chosen with a momentum

<sup>1</sup> [http://www.festvox.org/cm\\_u\\_arctic/](http://www.festvox.org/cm_u_arctic/)



**Fig. 10** Influence of the GA-SS approach on the average RMS error. We present the average RMS error over 20 synthesized sentences per each speaker. “SLT” is an American English female speaker, whereas “BDL” and “JMK” are American and Canadian English male speakers, respectively

of 0.3. More specifically, after 10 epochs, the momentum was increased to 0.9 and then the learning rate was halved regularly. The FF-DNN used in this work was implemented in the open source Merlin toolkit for speech synthesis [68] with some modifications. Besides, the training procedures were conducted on a high performance NVidia Titan X GPU. Weights and biases were prepared with small nonzero values, and optimized with stochastic gradient descent to minimize the mean squared error between its predictions and acoustic features of the training set.

### 3.3 Error measurement metrics

It is well-known that the efficient method for evaluating speech quality is typically done through subjective listening tests. However, there are various issues related with the use of subjective testing. It can be sometimes very expensive, time consuming, and hard to find a sufficient number of suitable volunteers [22, 51]. For that reason, it can often be useful in this work to run objective tests in addition to listening tests. Similarly, finding a meaningful objective metric is always a challenge in evaluating the performance of speech quality, similarity, and intelligibility. In fact, one metric possibly suitable for a few systems but not convenient for all. The reason for that may be returned to some factors which are influenced by the speed, complexity, or accuracy of the speech models. Speaker types and environmental conditions should also be taken into account when choosing these metrics. Therefore, a range of objective speech quality and intelligibility measures are considered to evaluate the quality of the proposed model. The results were averaged over the test utterances for each speaker. The following seven evaluation metrics were used:

- a) **Weighted Spectral Slope (WSS)** [30]: The algorithm first decomposes the frame signal into a set of frequency bands. The intensities within each critical band are measured. Then, a weighted distance between the measured slopes of the log-critical band spectra are computed

$$WSS = \frac{1}{N} \sum_{j=1}^N \left( \frac{\sum_{i=1}^K W_{i,j} (Y_{i,j} - X_{i,j})^2}{\sum_{i=1}^K W_{i,j}} \right) \tag{8}$$

where  $N$  is the number of frames in the utterance, and  $K$  is the number of sub-bands.  $W_{i,j}$ ,  $X_{i,j}$ , and  $Y_{i,j}$  denote the weight, the spectral slope of target speech signal, and the spectral slope of converted speech signals; respectively, at the  $i^{th}$  frequency band and  $j^{th}$  frame.

- b) **Log-Likelihood Ratio (LLR)** [51]: It is a distance measure that can be calculated from the linear prediction coefficients (LPC) vector of the target and converted speech. The segmental LLR is

$$LLR = \frac{1}{N} \sum_{i=1}^N \log \left( \frac{a_{y,i}^T R_{x,i} a_{y,i}}{a_{x,i}^T R_{x,i} a_{x,i}} \right) \tag{9}$$

where  $a_x$ ,  $a_y$ , and  $R_x$  are the LPC vector of the target signal frame, converted signal frame, and the autocorrelation matrix of the target speech signal, respectively.

- c) **Itakura-Saito (IS)** [25]: It is also a distance measure computed from the LPC vector

$$IS = \frac{1}{N} \sum_{i=1}^N \frac{\sigma_{x,i}^2}{\sigma_{y,i}^2} \left( \frac{a_{y,i}^T R_{x,i} a_{y,i}}{a_{x,i}^T R_{x,i} a_{x,i}} \right) + \log \left( \frac{\sigma_{y,i}^2}{\sigma_{x,i}^2} \right) - 1 \tag{10}$$

where  $\sigma_x^2$  and  $\sigma_y^2$  are the LPC all-pole gains of the target and converted signal frames, respectively.

- d) **Log Spectral Distortion (LSD)**: It can be defined as the square difference carried over the logarithm of the spectral envelopes of target  $X(f)$  and converted  $Y(f)$  speech signals at  $N$  frequency points

$$LSD = \sqrt{\frac{1}{N} \sum_{i=1}^N \text{mean} (\log X(f_i) - \log Y(f_i))^2} \tag{11}$$

- e) **Normalized Covariance Metric (NCM)** [37]: It is based on a Speech Transmission Index (STI) [55], which uses covariance coefficient  $r$  of the Hilbert envelope between the target and converted frame signal

$$NCM = \frac{1}{N} \sum_{j=1}^N \left( \frac{\sum_{i=1}^K W_{i,j} \cdot \log \frac{r_{i,j}^2}{1-r_{i,j}^2}}{\sum_{i=1}^K W_{i,j}} \right) \tag{12}$$

where  $W$  is the weight vector applied to the STI of  $K$  bands and can be found by the articulation index [4].

- f) **frequency-weighted segmental SNR (fwsNRseg)** [37]: Similarly to Equation (12),  $fwsNR_{seg}$  can be estimated by

$$fwsNR_{seg} = \frac{1}{N} \sum_{j=1}^N \left( \frac{\sum_{i=1}^K W_{i,j} \cdot \log \frac{X_{i,j}^2}{X_{i,j}^2 - Y_{i,j}^2}}{\sum_{i=1}^K W_{i,j}} \right) \tag{13}$$

where  $X_{i,j}^2$ ,  $Y_{i,j}^2$  are critical-band magnitude spectra in the  $j^{th}$  frequency band of the target and converted frame signals respectively,  $K$  is the number of bands,  $W$  is the weight vector defined in [4].

- g) **Mel-Cepstral Distortion (MCD)** [33]: It is based on the Euclidean distance between the target and converted frame vectors that describe the global spectral characteristics.

$$MCD = \frac{1}{N} \sum_{j=1}^N \sqrt{\sum_{i=1}^K (x_{i,j} - y_{i,j})^2} \quad (14)$$

where  $x$  and  $y$  are the  $i^{\text{th}}$  cepstral coefficients of the target and converted speech signals, respectively.

## 4 Evaluation results and discussion

The experimental evaluation has two main goals. First, it aims to evaluate the quality of the generated speech with respect to naturalness. The second goal is to evaluate how similar is the converted speech to the target speaker. Therefore, a reference (baseline) system with high quality performance is required to demonstrate the effectiveness and performance of the proposed methodology. Since the WORLD<sup>2</sup> vocoder [42] has a high-quality speech synthesis system for real-time applications and better than several high-quality vocoders (such as STRAIGHT) [40], we use it as our state-of-the-art baseline within SVC. We did not use WaveNet or Tacotron based neural vocoders as a baseline in the experiment because our proposed vocoder is a source-filter based system, not an end-to-end acoustic model. For the VC experiments using the WORLD vocoder, we used the same FF-DNN architecture as for the proposed vocoder (see Sec. 2.2 and 3.2). We have synthesized 20 utterances for each of the speaker pair conversions, which means that 80 sentences are available for evaluation.

### 4.1 Objective evaluation

Here, we show the results for the error metrics presented in subsection 3.3. For all empirical measures, a calculation is done frame-by-frame, and a lower value indicates better performance except for the fwSNRseg measure (higher value is better). The results were averaged, and the best value in each column of Table 1 is bold faced.

It is interesting to emphasize that the findings in Table 1 showed that the baseline does not meet the performance of our proposed model. That, in other words, the results reported in Table 1 strongly support the use of the proposed vocoder for SVC. In particular, the fwSNRseg between converted and target speech frames using the proposed method with continuous vocoder are higher than those using the baseline method. Nevertheless, the WORLD vocoder is shown to be better only for the SLT-to-JMK speaker conversion.

The comparison of the spectral envelope of one speech frame converted by the proposed method is given in Fig. 11. The converted spectral envelope is plotted along with the source and the preferred target. It may be observed that the converted spectral envelope is more similar in general to the target one than the source one. Even though, these two trajectories seem similar, they are moderately smoothed compared with the target one; that can affect the quality of the converted speech. It can also be seen in Fig. 12 that the converted contF0

<sup>2</sup> <https://github.com/mmorise/World>

**Table 1** Average scores on converted speech signal per each of the speaker pairs conversion

Error metrics	Model	SLT-to-BDL	BDL-to-SLT	SLT-to-JMK	JMK-to-SLT
MCD	Reference	5.624	5.355	5.856	5.765
	Proposed	<b>5.609</b>	<b>5.341</b>	<b>5.846</b>	<b>5.754</b>
fwSNR <sub>seg</sub>	Reference	1.660	1.119	<b>2.162</b>	0.558
	Proposed	<b>3.072</b>	<b>1.873</b>	1.970	<b>1.312</b>
LSD	Reference	2.423	2.208	2.506	2.557
	Proposed	<b>2.214</b>	<b>2.107</b>	<b>2.368</b>	<b>2.401</b>
IS	Reference	33.005	24.887	33.060	39.418
	Proposed	<b>15.183</b>	<b>21.212</b>	<b>13.973</b>	<b>29.137</b>
WSS	Reference	8.842	16.299	8.068	17.310
	Proposed	<b>7.723</b>	<b>13.683</b>	<b>7.783</b>	<b>14.046</b>
LLR	Reference	1.718	1.724	1.610	1.744
	Proposed	<b>1.451</b>	<b>1.581</b>	<b>1.442</b>	<b>1.640</b>
NCM	Reference	0.103	0.102	0.024	0.030
	Proposed	<b>0.115</b>	<b>0.124</b>	<b>0.028</b>	<b>0.035</b>

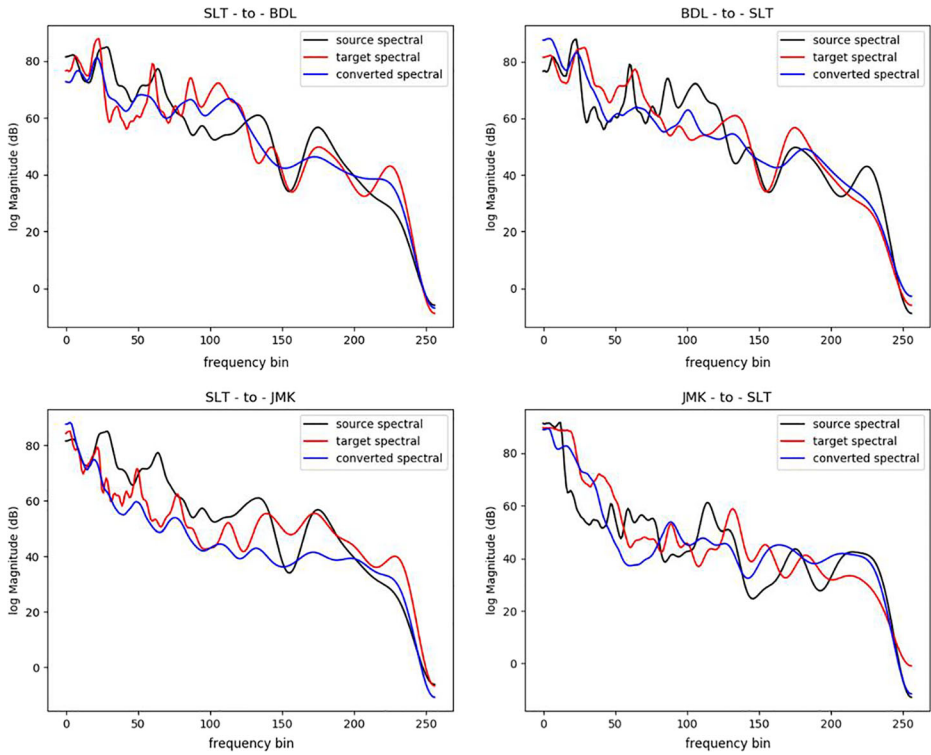
trajectories generated from proposed method follow the same shape of the target confirming the similarity between them and can provide better F0 predictions. Similarly, when looking at Fig. 13, it makes apparent that the proposed framework produces converted speech with MVF more similar to the target trajectories rather than to the source ones.

As a result, these experiments show that the proposed model with continuous vocoder is competitive for the SVC task, and superior to the reference WORLD model.

## 4.2 Subjective evaluation

To demonstrate the efficiency of our proposed model, we conducted two different perceptual listening tests. First, in order to evaluate the similarity of the converted speech to a reference target voice (which was the natural voice), we performed a web-based MUSHRA-like (Multi-Stimulus test with Hidden Reference and Anchor) listening test [26]. The advantage of MUSHRA is that it enables evaluation of multiple samples in a single trial without breaking the task into many pairwise comparisons, and it is a standard method for speech synthesis evaluations. Within the MUSHRA test we compared four variants of the sentences: 1) Source, 2) Target, 3) Converted speech using the high-quality baseline (WORLD) vocoder, 4) Converted speech using the proposed (Continuous) vocoder. Listeners were asked to assess which variant was more similar to the natural reference (i.e., the target speaker), without considering the speech quality, from 0 (highly not similar) to 100 (highly similar). From the testing set of Sec. 4.1, 12 utterances were randomly chosen and presented in a randomized order (different for each participant). Altogether, 48 utterances were included in the MUSHRA test (4 types  $\times$  12 sentences).

Second, in order to evaluate the overall quality and identity of the synthesized speech from both proposed and baseline systems, a Mean Opinion Score (MOS) test was carried out. In the MOS test we compared three variants of the sentences: 1) Target, 2) Converted speech using the baseline (WORLD) vocoder, and 3) Converted speech using the proposed (Continuous) vocoder. The listeners had to rate the naturalness of each stimulus, from 0 (highly unnatural) to 100 (highly natural). Similarly, the same 12 sentences were used as in the MUSHRA test. Altogether, 36 utterances were included in the MOS test (3 types  $\times$  12 sentences).



**Fig. 11** Example of one shorter segment /e/ from the natural source, target, and converted spectral envelopes using proposed method. Sentence: “Gad, your letter came just in time”

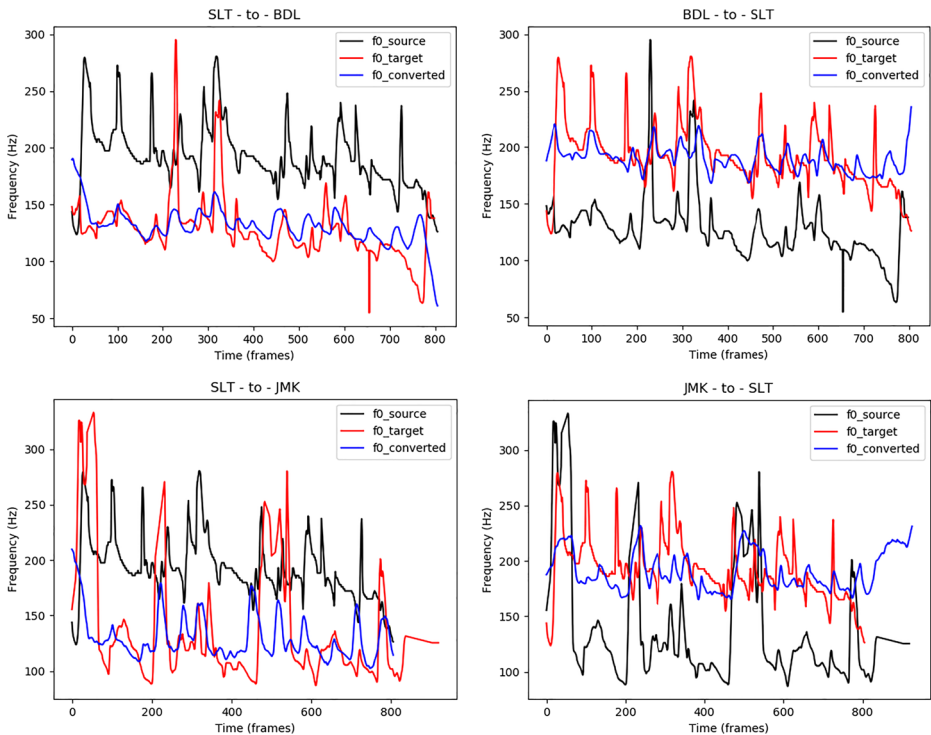
Before the test, listeners were asked to listen to an example from the male speaker to adjust the volume. Nineteen participants between the age of 23–40 (mean age: 30 years) were asked to conduct the online listening test. 12 of them were males and 7 were females. On average, the MUSHRA test took 13 min, while the MOS test was 12 min long. The listening tests samples can be found online.<sup>3</sup>

The MUSHRA similarity scores of the listening test are presented in Fig. 14. It can be seen that both systems achieve almost similar performance to the target voice across all gender combinations. This means that our proposed model has successfully converted the source voice to the target voice on cross-gender cases. In case of SLT-to-BDL conversion, the difference between the baseline and the proposed systems is statistically significant (Mann-Whitney-Wilcoxon ranksum test, with a 95% confidence level), while the other differences between the baseline and proposed are not significant.

Additionally, Fig. 15 shows the results of the MOS test. We can see that both the baseline and proposed systems achieved low naturalness scores compared to the target speaker, showing that the listeners clearly differentiated the utterances resulting from the voice conversion. It can be also found that the listeners preferred the baseline system compared to the proposed one. However, none of these differences are statistically significant (Mann-Whitney-Wilcoxon ranksum test, with a 95% confidence level).

<sup>3</sup> <http://smartlab.tmit.bme.hu/vc2019>





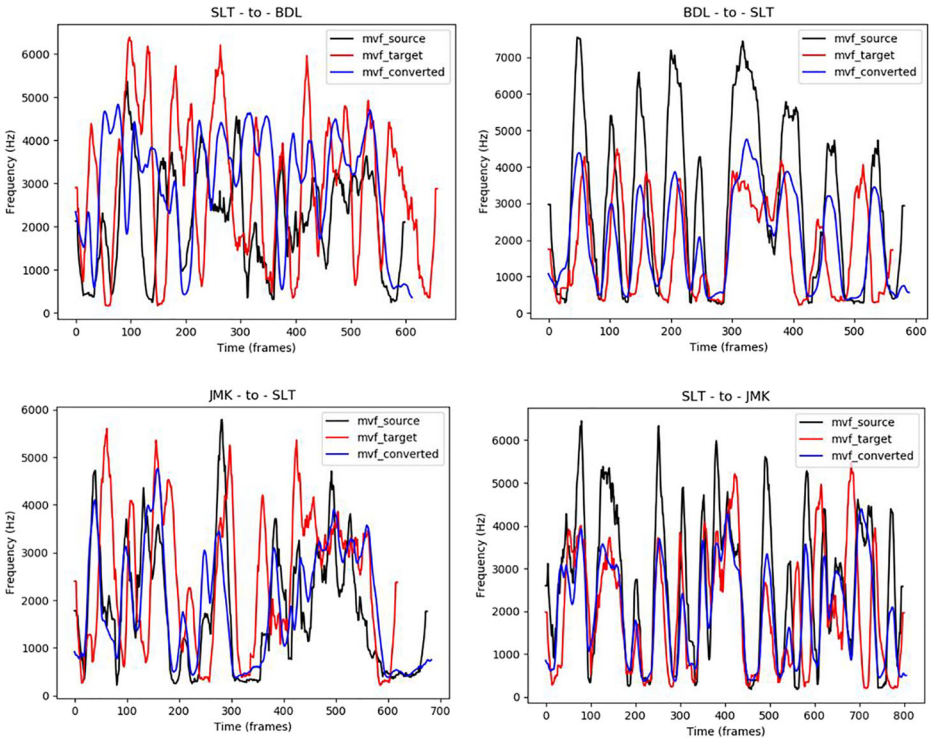
**Fig. 12** Example of the natural source, target, and converted contF0 trajectories using proposed method. Sentence: “From that moment his friendship for Belize turns to hatred and jealousy”

As the final result of the listening tests investigating similarity to the target speaker and overall quality, we can conclude that the proposed continuous vocoder within the SVC framework performed well, because it is as good as the voice conversion using the WORLD vocoder.

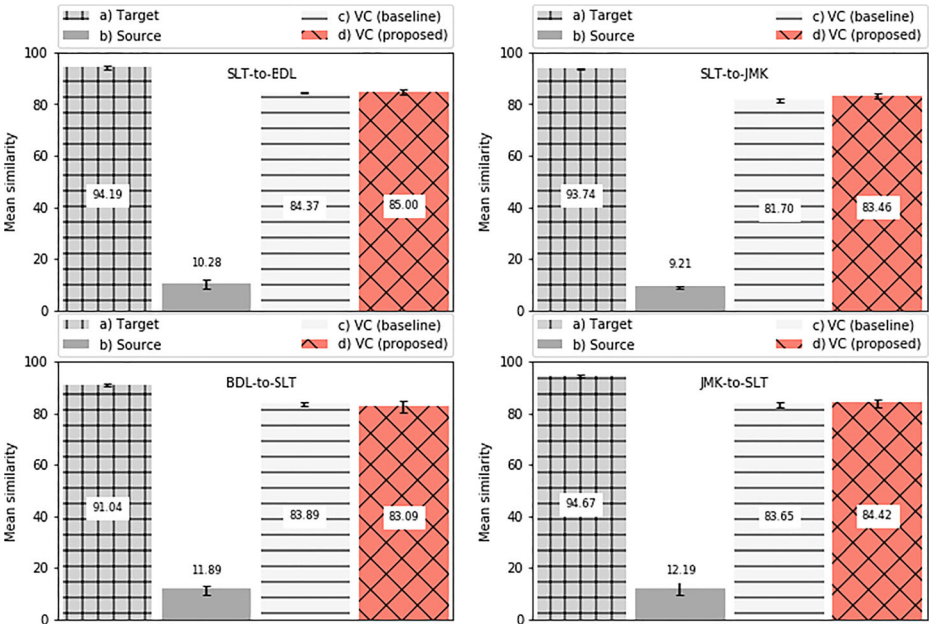
## 5 Conclusions

In this paper, we proposed a new approach to statistical voice conversion using a feed-forward deep neural network. The main idea was to integrate the continuous vocoder into the SVC framework, which provides an advanced model of the excitation signal, by converting its contF0, MVF, and spectral features within a statistical conversion function. The advantage of this vocoder is that it does not require to have a voiced/unvoiced decision, which means that the alignment error will be avoided in SVC between voiced and unvoiced segments. Therefore, its simplicity and flexibility allows us to easily construct a voice conversion framework using a FF-DNN.

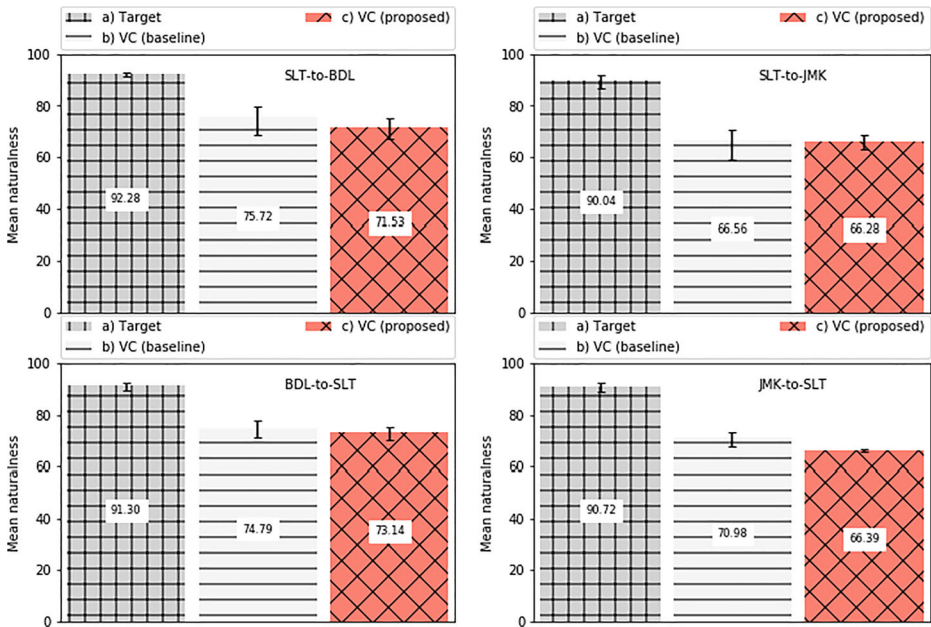
Using a variety of measurements, the performance strengths and weaknesses of the proposed method for different speakers were highlighted. From the objective experiments, the performance of the proposed system (using the continuous vocoder) was superior in most cases to that of the reference system (using the WORLD vocoder). Moreover, two listening tests have been performed to evaluate the effectiveness of the proposed method. The similarity test showed that the reference and proposed systems are both similar to the target speaker. This also confirms our findings, that are reported in the objective evaluations. Significant



**Fig. 13** Example of the natural source, target, and converted MVF contours using the proposed method. Sentence: “Gregson shoved back his chair and rose to his feet”



**Fig. 14** MUSHRA scores for the similarity question. Higher value means larger similarity to the target speaker. Errorbars show the bootstrapped 95% confidence intervals



**Fig. 15** MOS scores for the naturalness question. Higher value means better overall quality. Errorbars show the bootstrapped 95% confidence intervals

differences were not found compared to the reference system during the quality (MOS) test. This means that the proposed approach is capable of converting speech with higher naturalness and perceptual speech intelligibility.

Plans of future research involve first to add a Harmonics-to-Noise Ratio as a new parameter to the analysis, statistical learning and synthesis steps in order to further reduce the buzziness caused by vocoding. Secondly, it would be interesting to investigate the effectiveness of applying a mixture density recurrent network by using a bi-directional long-short memory (Bi-LSTM) based SVC to further improve the perceptual quality of the converted speech.

**Acknowledgements** The research was partly supported by the AI4EU project and by the National Research, Development and Innovation Office of Hungary (FK 124584 and PD 127915 grants). The Titan X GPU used was donated by NVIDIA Corporation. We would like to thank the subjects for participating in the listening test.

**Funding Information** Open access funding provided by Budapest University of Technology and Economics (BME).

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- Aihara R, Takiguchi T, Ariki Y (2014) Individuality-preserving voice conversion for articulation disorders using dictionary selective non-negative matrix factorization. In: Proceedings of SLPAT, p 29–37

2. Al-Radhi MS, Csapó TG, Németh G (2017) Continuous vocoder in feed-forward deep neural network based speech synthesis. In: Proceedings of the digital speech and image processing. Serbia
3. 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 Interspeech. Stockholm, p 434–438
4. ANSI (1997) Methods for the calculation of the speech intelligibility index. American National Standards Institute, ANSI Standard S3.5
5. Chen LH, Ling ZH, Liu LJ, Dai LR (2014) Voice conversion using deep neural networks with layer-wise generative training. *IEEE Trans Audio Speech Lang Process* 22(12):1859–1872. <https://doi.org/10.1109/TASLP.2014.2353991>
6. Childers D (1995) Glottal source modeling for voice conversion. *Speech Comm* 16(2):127–138. [https://doi.org/10.1016/0167-6393\(94\)00050-K](https://doi.org/10.1016/0167-6393(94)00050-K)
7. Childers DG, Yegnanarayana B, Wu K (1985) Voice conversion: factors responsible for quality. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, Tampa, FL, USA p 748–751. <https://doi.org/10.1109/ICASSP.1985.1168479>
8. Childers DG, Wu K, Hicks DM, Yegnanarayana B (1989) Voice conversion. *Speech Comm* 8(2):147–158. [https://doi.org/10.1016/0167-6393\(89\)90041-1](https://doi.org/10.1016/0167-6393(89)90041-1)
9. Csapó TG, Németh G, Cernak M, Garner PN (2016) Modeling unvoiced sounds in statistical parametric speech synthesis with a continuous vocoder. In: Proceedings of 24th European Signal Processing Conference (EUSIPCO), Budapest, p 1338–1342
10. Desai S, Raghavendra EV, Yegnanarayana B, Black AW, Prahallad K (2009) Voice conversion using artificial neural networks. In: Proceedings of ICASSP, p 3893–3896
11. Doi H, Toda T, Nakamura K, Saruwatari H, Shikano K (2014) Alaryngeal speech enhancement based on one-to-many eigenvoice conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22(1):172–183. <https://doi.org/10.1109/TASLP.2013.2286917>
12. 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>
13. 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>
14. Drugman T, Wilfart G, Dutoit T (2009) Eigenresiduals for improved parametric speech synthesis. In: Proc. EUSIPCO. Glasgow, p 2176–2180
15. Drugman T, Thomas M, Gudnason J, Naylor P, Dutoit T (2012) Detection of glottal closure instants from speech signals: a quantitative review. *IEEE Trans Audio Speech Lang Process* 20(3):994–1006. <https://doi.org/10.1109/TASL.2011.2170835>
16. Dutoit T (1997) High-quality text-to-speech synthesis: an overview. *J Electr Electron Eng Aust* 17:25–36
17. Erro D, Sainz I, Navas E, Hernaez I (2014) Harmonics plus noise model based vocoder for statistical parametric speech synthesis. *IEEE Journal of Selected Topics in Signal Processing* 8(2):184–194. <https://doi.org/10.1109/JSTSP.2013.2283471>
18. 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>
19. Hashimoto K, Oura K, Nankaku Y, Tokuda K (2015) The effect of neural networks in statistical parametric speech synthesis. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), p 4455–4459
20. Helander E, Silen H, Virtanen T, Gabbouj M (2012) Voice conversion using dynamic kernel partial least squares regression. *IEEE Trans Audio Speech Lang Process* 20(3):806–817. <https://doi.org/10.1109/TASL.2011.2165944>
21. Hideyuki M, Masanobu A (1995) Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt. *Speech Comm* 16(2):153–164. [https://doi.org/10.1016/0167-6393\(94\)00052-C](https://doi.org/10.1016/0167-6393(94)00052-C)
22. Hu Y, Loizou PC (2008) Evaluation of objective quality measures for speech enhancement. *IEEE Trans Audio Speech Lang Process* 16(1):229–238. <https://doi.org/10.1109/TASL.2007.911054>
23. Hu Q, Richmond K, Yamagishi J, Latorre J (2013) An experimental comparison of multiple vocoder types. In: Proceedings of the ISCA SSW8, p 155–160
24. 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>
25. Itakura F, Saito S (1968) An analysis-synthesis telephony based on the maximum-likelihood method. In: Proc Int Congr Acoust, Japan, p C17–C20

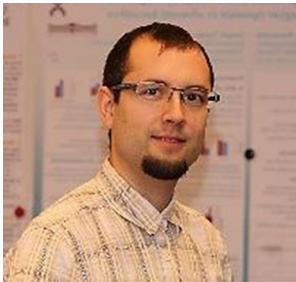
26. ITU-R Recommendation BS.1534 (2001) Method for the subjective assessment of intermediate audio quality
27. Kai Y, Steve Y (2011) Continuous F0 modelling for HMM based statistical parametric speech synthesis. *IEEE Trans Audio Speech Lang Process* 19(5):1071–1079. <https://doi.org/10.1109/TASL.2010.2076805>
28. 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
29. Kenmochi H (2012) Singing synthesis as a new musical instrument. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, p 5385–5388
30. Klatt D (1982) Prediction of perceived phonetic distance from critical band spectra: a first step. In: *IEEE proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. New York, p 1278–1281
31. Kobayashi K, Hayashi T, Tamamori A, Toda T (2017) Statistical voice conversion with WaveNet-based waveform generation. In: *Proceedings of Interspeech*, Stockholm, Sweden, p 1138–1142
32. Kominek J, Black AW (2003) CMU ARCTIC databases for speech synthesis. Carnegie Mellon University
33. Kubichek RF (1993) Mel-cepstral distance measure for objective speech quality assessment. In: *Proceedings of IEEE Pacific rim conference on communications, computers and signal processing*, p 125–128
34. Lenarczyk M (2014) Parametric speech coding framework for voice conversion based on mixed excitation model. In: *International Conference on Text, Speech, and Dialogue*. Lecture Notes in Computer Science. Springer, Cham, 8655, p 507–514
35. Lifang W, Linghua Z (2012) A voice conversion system based on the harmonic plus noise excitation and Gaussian mixture model. In: *International Conference on Instrumentation, Measurement, Computer, Communication and Control*, p 1575–1578
36. Ling ZH, Deng L, Yu D (2013) Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis. *IEEE Trans Audio Speech Lang Process* 21(10): 2129–2139. <https://doi.org/10.1109/TASL.2013.2269291>
37. Ma J, Hu Y, Loizou P (2009) Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *The Journal of the Acoustical Society of America* 125(5):3387–3405. <https://doi.org/10.1121/1.3097493>
38. Mohammadi SH, Kain A (2014) Voice conversion using deep neural networks with speaker-independent pre-training. In: *IEEE Spoken Language Technology Workshop (SLT)*, NV, USA, p 19–23. <https://doi.org/10.1109/SLT.2014.7078543>
39. 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>
40. Morise M, Watanabe Y (2018) Sound quality comparison among high-quality vocoders by using re-synthesized speech. *Acoust Sci Technol* 39(3):263–265. <https://doi.org/10.1250/ast.39.263>
41. Morise M, Kawahara H, Nishiura T (2010) Rapid f0 estimation for high SNR speech based on fundamental component extraction. *IEICE Trans Inf Syst* J93-D(2):109–117
42. 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
43. 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>
44. Nakamura K, Toda T, Saruwatari H, Shikano K (2010) The use of air-pressure sensor in electrolaryngeal speech enhancement based on statistical voice conversion. In: *Proceedings of Interspeech*, Makuhari, Japan, p 1628–1631
45. Nakashika T, Takashima T, Takiguchi R, Ariki Y (2013) Voice conversion in high-order eigen space using deep belief nets. In: *Proceedings of Interspeech*, p 369–372
46. Nakashika T, Takiguchi T, Ariki Y (2014) High-order sequence modeling using speaker dependent recurrent temporal restricted Boltzmann machines for voice conversion. In: *Proceedings of the annual conference of the international speech communication association*, p 2278–2282
47. New TL, Dong M, Chan P, Wang X, Ma B, Li H (2010) Voice conversion: from spoken vowels to singing vowels. In: *Proc IEEE Int Conf Multimedia Expo (ICME)*, Suntec City, Singapore, p 1421–1426. <https://doi.org/10.1109/ICME.2010.5582961>
48. 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>
49. Nose T, Kobayashi T (2011) Speaker-independent HMM-based voice conversion using adaptive quantization of the fundamental frequency. *Speech Comm* 53(7):973–985. <https://doi.org/10.1016/j.specom.2011.05.001>

50. Qian Y, Fan Y, Hu W, Soong FK (2014) On the training aspects of deep neural network (DNN) for parametric TTS synthesis. In: Proc IEEE Int Conf on Acoustics, Speech, and Signal Processing (ICASSP), p 3829–3833
51. Quackenbush S, Barnwell T, Clements M (1988) Objective measures of speech quality. Prentice-Hall, Englewood Cliffs
52. Sakoe H, Chiba S (1978) Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans Acoust Speech Signal Process 26(1):43–49. <https://doi.org/10.1109/TASSP.1978.1163055>
53. Sisman B, Li H (2018) Wavelet analysis of speaker dependent and independent prosody for voice conversion. In: Proceedings of Interspeech, Hyderabad, p 52–56
54. Sisman B, Zhang M, Sakti S, Li H, Nakamura S (2018) Adaptive WaveNet vocoder for residual compensation in GAN-based voice conversion. In: IEEE spoken language technology workshop (SLT). Athens
55. Steeneken H, Houtgast T (1980) A physical method for measuring speech-transmission quality. The Journal of the Acoustical Society of America 67(1):318–326. <https://doi.org/10.1121/1.384464>
56. Stylianou Y (2001) Applying the harmonic plus noise model in concatenative speech synthesis. IEEE Trans Speech Audio Process 9:21–29. <https://doi.org/10.1109/89.890068>
57. Takamichi S, Toda T, Black AW, Neubig G, Sakti S, Nakamura S (2016) Postfilters to modify the modulation spectrum for statistical parametric speech synthesis. IEEE Trans Audio Speech Lang Process 24(4):755–767. <https://doi.org/10.1109/TASLP.2016.2522655>
58. Tokuda K, Kobayashi T, Masuko T, Imai S (1994) Mel-generalized cepstral analysis - a unified approach to speech spectral estimation. In: Proc. of the ICSLP, p 1043–1046
59. Tomoki T, Black AW, Keiichi T (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>
60. Tóth BP, Csapó TG (2016) Continuous fundamental frequency prediction with deep neural networks. In: Proc. of the European signal processing conference (EUSIPCO). Budapest
61. Valentini-Botinhao C, Wu Z, King S (2015) Towards minimum perceptual error training for DNN-based speech synthesis In: Proceedings of Interspeech, p 869–873
62. Wang Y, et al. (2017) Tacotron: towards end-to-end speech synthesis. In: Proceedings of Interspeech. Stockholm, p 4006–4010
63. Wu Z, Virtanen T, Chng ES, Li H (2014) Exemplar-based sparse representation with residual compensation for voice conversion. IEEE/ACM Transactions on Audio, Speech and Language Processing 22(1):1506–1521 <https://doi.org/10.1109/TASLP.2014.2333242>
64. Wu Z, Valentini-Botinhao C, Watts O, King S (2015) Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), p 4460–4464
65. Yang L, Philippos CL (2008) A geometric approach to spectral subtraction. Speech Comm 50(6):453–466. <https://doi.org/10.1016/j.specom.2008.01.003>
66. Zen H, Senior A, Schuster M (2013) Statistical parametric speech synthesis using deep neural networks. In: Proc IEEE Int Conf on Acoustics, Speech, and Signal Processing g (ICASSP), p 7962–7966
67. Zhang Q, Soong F, Qian Y, Yan Z, Pan J, Yan Y (2010) Improved modeling for F0 generation and V/U decision in HMM-based TTS. In: Proc of the IEEE Int Conf Acoustics, Speech and Signal Processing, Dallas
68. Zhizheng W, Watts O, King S (2016) Merlin: an open source neural network speech synthesis system. In Proceeding 9th ISCA speech synthesis workshop (SSW9). Sunnyvale

**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.

## Affiliations

**Mohammed Salah Al-Radhi<sup>1</sup> · Tamás Gábor Csapó<sup>1,2</sup> · Géza Németh<sup>1</sup>**

<sup>1</sup> Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary

<sup>2</sup> MTA-ELTE Lendület Lingual Articulation Research Group, Budapest, Hungary