Multimedia Tools and Applications (2023) 82:15635–15649
https://doi.org/10.1007/s11042-022-14005-5

Investigations on speaker adaptation using a continuous vocoder within recurrent neural network based text-to-speech synthesis

Ali Raheem Mandeel[1] · **Mohammed Salah Al-Radhi**[1] · **Tamás Gábor Csapó**[1]

Received: 22 October 2021 / Revised: 28 July 2022 / Accepted: 19 September 2022 /
Published online: 22 October 2022

Abstract

This paper presents an investigation of speaker adaptation using a continuous vocoder for parametric text-to-speech (TTS) synthesis. In purposes that demand low computational complexity, conventional vocoder-based statistical parametric speech synthesis can be preferable. While capable of remarkable naturalness, recent neural vocoders nonetheless fall short of the criteria for real-time synthesis. We investigate our former continuous vocoder, in which the excitation is characterized employing two one-dimensional parameters: Maximum Voiced Frequency and continuous fundamental frequency (F0). We show that an average voice can be trained for deep neural network-based TTS utilizing data from nine English speakers. We did speaker adaptation experiments for each target speaker with 400 utterances (approximately 14 minutes). We showed an apparent enhancement in the quality and naturalness of synthesized speech compared to our previous work by utilizing the recurrent neural network topologies. According to the objective studies (Mel-Cepstral Distortion and F0 correlation), the quality of speaker adaptation using Continuous Vocoder-based DNN-TTS is slightly better than the WORLD Vocoder-based baseline. The subjective MUSHRA-like test results also showed that our speaker adaptation technique is almost as natural as the WORLD vocoder using Gated Recurrent Unit and Long Short Term Memory networks. The proposed vocoder, being capable of real-time synthesis, can be used for applications which need fast synthesis speed.

Keywords Speech synthesis · RNN · TTS · Continuous vocoder

✉ Ali Raheem Mandeel
aliraheem.mandeel@edu.bme.hu

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

Tamás Gábor Csapó
csapot@tmit.bme.hu

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

1 Introduction

Text-to-speech (TTS) synthesis, a field of speech processing, converts the written text into speech. It is used in many application and industrial fields such as smartphones, railway announcements, robotics, healthcare, and education [17, 37]. It offers many benefits for students with a learning disability [5]. It can be used to overcome time consumed by reading via audiobooks using a TTS. It can also enhance literacy skills, improve reading comprehension, the ability to recall information, and the pronunciation improvement [38, 40]. It also makes it easy for anyone else searching for more convenient access to digital materials. TTS health-assistance technologies would significantly enhance people's health and lower their healthcare expenditures [22, 23].

One type of text-to-speech synthesis employs statistical parametric synthesis (SPSS), with two primary machine learning techniques: deep neural networks (DNN) and Hidden Markov models (HMM). DNNs have become the prevalent acoustic models for text-to-speech synthesis. It is used to generate high-level data representations.

Furthermore, a considerable improvement in voice quality may be accomplished with the availability of multi-task learning [45, 49]. In the parametric model, the speech signal is processed into parameters indicating fundamental frequency (F0) and speech spectrum, subsequently, input into a machine learning system. After the statistical model has been learned on the training data, the parameter sequences are rebuilt into speech signals utilizing reconstruction techniques (excitation models, vocoders).

SPSS naturalness can be improved through further developing vocoding approaches. Although neural vocoding systems can produce speech similar to natural speech (e.g., WaveNet [41]), and WaveGlow [32], their high computational complexity remains problematic. Additionally, they necessitate vast data and processing resources, making it challenging to train for real-time applications, particularly in embedded environments. SPSS, which uses vocoders to model high-quality synthetic speech, is an efficient and more easily adaptable solution. Besides, controllability options make vocoder-based SPSS preferable over neural vocoders.

Deep learning algorithms perform exceptionally well when training on a large enough dataset. Subsequently, this technique means that to train a model for speech synthesis, we need a lot of speech from a target speaker (e.g., 60 minutes). Eventually, it limits the technology's capacity to scale to a variety of voices. In TTS, the simple fine-tuning creates a high-quality model when the data from the target speaker is adequate [20]. In this case, we can restrict the tuning to a particular part alternately to the whole system to avoid overfitting [9, 26] Adaptation methods may be used to produce new voices using only a small amount of adaptation data from a target speaker [27]. Because of the potential performance of speaker adaptation, SPSS has a significant advantage over unit-selection speech synthesis in terms of changing speaker attributes, emotions, and speaking styles [25]. Recent neural vocoders can produce incredibly natural speech, yet they usually fall short of real-time synthesis requirements [30, 34].

We proposed in our previous work [28] a speaker adaptation system based on a continuous vocoder and a Feed-forward deep neural network (FFNN) text-to-speech synthesis system. In this paper, we investigated the recurrent neural networks (RNNs) such as Long Short Term Memory networks (LSTM) and Gated Recurrent Unit (GRU) in both WORLD vocoder (baseline system) and continuous vocoder. As well, on the basis of data from nine speakers, we show that an average voice can be trained for DNN-TTS and that speaker adaptability is possible with 400 utterances (about 14 minutes).

The remaining part of the paper is structured as follows. Section 2 summarizes the speech signal structure. Section 3 comprises a summary of relevant scientific publications, including innovative approaches in text-to-speech synthesis based on speaker adaptation. In Section 4, the continuous vocoder is introduced in detail and compared to other vocoders. The design of the system, tools, and dataset are described in Section 5. The results are then explained in (Section 6). Finally, in Section 7, the conclusion is given.

2 Speech signal structure

A concept is initially generated in the speaker's brain at the linguistic level of communication before being translated into words, phrases, and sentences. The brain makes electrical signals that move at the nerves at the physiological level. The vocal cords and tract muscles are triggered by these electrical signals. Changes in vocal tract pressure are caused by the movement of the vocal cords and vocal tract, and the lips produce the sounds [33].

The vibration of vocal cords creates the voiced sounds, and the rate of vibration of vocal cords is referred to as a fundamental frequency (F0). The perceptual correlate of F0 is named pitch. The formant frequencies generally are the energy attention at higher frequency ranges.

In the time domain, the speech waveform represents time on the linear horizontal axis and amplitude on the vertical axis. The primary levels of timing information in the speech acoustic are envelope, periodicity, and fine structure (the top part of Fig. 1). The envelope shows the slow outline of the overall changes in the intensity over the signal. The repeating peaks at a regular periodic rate mean we have voiced sounds. The fine structure possesses all high frequencies ups and downs in the waveform corresponding to the highest frequency changes. In the frequency spectrum, the frequency is on the horizontal axis and the amplitude on the vertical axis. Finally, the spectrogram shows the frequency content change over time the bottom part of (Fig. 1).

3 Related work

In addition to improving the naturalness of synthesized speech, synthesizing new voices with a small amount of data is an enthusiastic research topic. Numerous approaches are proposed to address this concern. They all share the same basic principle: use a large corpus to repay a target speaker's lack of data. Because DNNs have a high number of parameters and the parameters of the DNN model cannot be interpreted as directly as the parameters of the HMM model, DNN-based TTS adaptation is more complicated than HMM-based adaptation [7].

One way for speaker adaptation is to transmit speaker codes to the main network, which can then adjust the layer weights. These speaker codes can be learned or i-vectors. A supervised method of learning speaker codes for DNN-HMM systems' speaker adaptation was suggested by [48]. These algorithms learn a bias for the sigmoid nonlinearities at the product of fully connected layers employing speaker codes. Besides, a neural fusion architecture of a unit concatenation approach and SPSS to enhance the similarity of the synthesized speech to the target speaker was proposed [11]. Despite that, the naturalness has not been upgraded.

In addition, auxiliary speaker embedding-based DNN adaptation strategies encode speaker-dependent (SD) properties in a compact vector representation in additional speaker embedding-based approaches. SD embedding vectors are employed as supplemental input

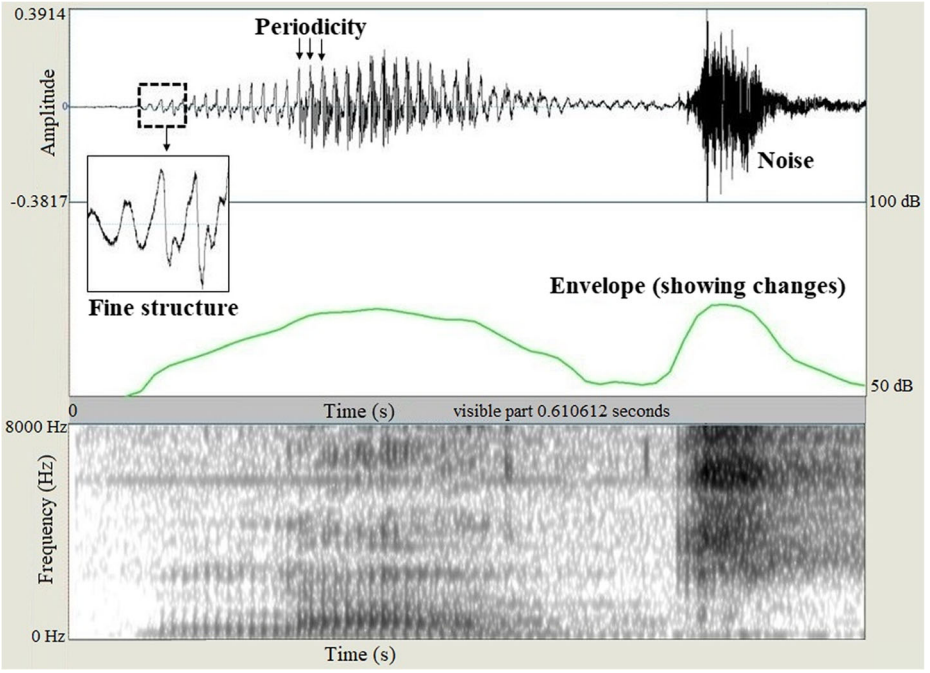

Fig. 1 The speech waveform and spectrogram info. of the word "nudge" (drawn using Praat software)

features for DNNs, allowing for more effortless speaker adaptability during system training and evaluation. The SD auxiliary feature estimation is done separately from the other components of the recognition system. An instance of i-vectors were learned from universal background models based on the Gaussian mixture model was given by [36]. D-vectors have been employed to transmit the speaker-specific data in adaptation [6]. A DNN speaker adaptation framework based on full Bayesian learning to simulate speaker-dependent (SD) parameter uncertainty with limited speaker-specific adaptation data was offered by [47].

Some speaker adaptation approaches are model-based adaptation techniques (e.g., the learning hidden unit contributions (LHUC)). The amount of speaker-specific adaptation data effectively influences the modeling subdivisions of the SD parameters in this design. When trained on a large quantity of speech data with multiple speakers, the LHUC technique posits that a neural network's hidden representations are base vectors learned to interpret the acoustic space of several speakers [44].

Zero-shot multi-speaker TTS (ZS-TTS) employs only a few seconds of speech dataset to create synthesized voices for target speakers. SC-GlowTTS is a novel model (GlowTTS and HiFi-GAN) to enhance the likeness to the target speaker using ZS-TTS [8]. The all pass warp (APW) has been integrated into neural network frameworks and investigated in ZS-TTS [35]. Their research shows that APW improves multi-speaker model generalizability and speaker similarity to the target speaker. Better speaker similarity is achieved by a multi-model speaker adaptation system (an unsupervised speech representation module and Tacotron2) than by a single-model (Tacotron2) [50]. In addition, intriguing research

develops a multi-model speaker adaptation, the Hieratron model framework, to control the prosody and timbre and synthesize the speech from a limited dataset [14].

Further, the few-shot adaptation technique modifies the network's parameters with the target speaker's speech. Employing two special encoders (fine-grained and coarse-grained) for cloning unseen target speakers with a few data samples has increased the similarity and quality of synthesized speech [12]. AdaSpeech 3 is an adaptive TTS model fine-tuning a trained reading style TTS model for spontaneous speech [24]. Using a small amount of target speech data, the model predicts filled pauses and changes in rhythm.

Hybrid deep learning (HDL) may represent the future direction of deep learning. It combines many artificial neural network concepts such as convolutional neural network (CNN), RNN, and Autocoder in one model. We previously tested our Continuous vocoder to build a deep learning model based on TTS by applying a Hybrid model of Bi-LSTM and conventional Uni-RNN [4]. In future research, we believe it is worth investigating the speaker adaptation with the Continuous vocoder using a hybrid model.

The fine-tuning approach belongs to the area of transfer learning, in which information gained from one activity may be applied to another. It was used in speech synthesis to train an average voice model (AVM) on a large database of speakers before applying the same network to a new speaker with fewer data. This technique was found to exceed the synthesized speech of LHUC and an HMM-based adaptation method [39]. Due to the lack of conducted research in speaker adaptation using the fine-tuning approach based-TTS, we used in the current study a novel technique utilization low complexity of continuous vocoder with a small database of target speakers to synthesize speech.

The vocoder is an essential aspect of TTS. It is responsible for translating human voices to vocoder parameters and vice versa. It is both an analyzer and a synthesizer. It analyzes speech by converting the waveform into a set of parameters that reflect the excitation signal of the vocal folds and then modifying the excitation signal with a vocal-tract filter transfer tool. Oppositely, it uses the parameters to reconstruct the original voice stream. The parametric representation allows for statistical modeling of speech as well as manipulation of speech to improve intelligibility. STRAIGHT, WORLD, Mel - generalised cepstral vocoder, adaptive harmonic model, Glottal vocoder, and Harmonic model are a few examples of vocoders [18].

4 Continuous vocoder

During the analysis stage of the Continuous vocoder [13], the fundamental frequency (contF0), is determined on the input waveforms. The maximum voiced frequency (MVF) parameter is then calculated from the speech signal. As the next step, the voice signal is subjected to 24-order Mel-Generalized Cepstral analysis (MGC) using gamma $= -1/3$ and alpha $= 0.42$. In all phases of the analysis, the frameshift is 5 ms. The Glottal Closure Instant (GCI) algorithm is used to find the glottal period restrictions of particular periods in the voiced parts of the inverse filtered residual (see Fig. 2).

After that in the synthesis stage, depending on the contF0, voiced excitation is made up of principal component analysis (PCA) residuals that overlap-added pitch simultaneously. Subsequently, at the frequency specified by the MVF parameter, this voiced excitation is low-pass filtered frame by frame. White noise is employed at frequencies greater than MVF's value. Then, both voiced and unvoiced excitation are combined. Applying a time

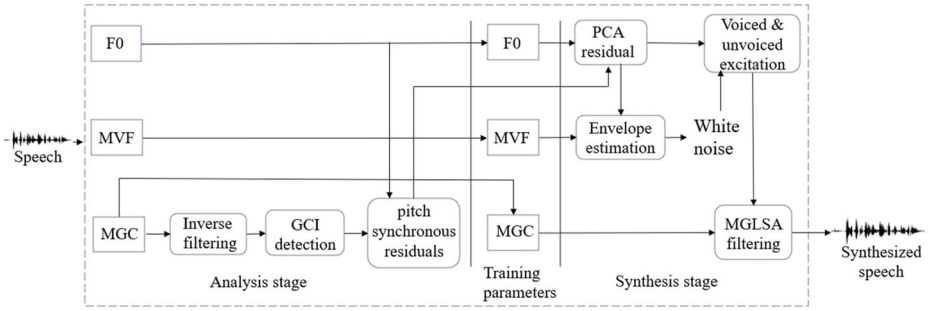

Fig. 2 General framework of the continuous vocoder

domain envelope to the unvoiced segments is presented as a means of modeling unvoiced sounds to eliminate any remaining buzziness and reduce the noise part [1, 3]. Consequently, to synthesize speech from the excitation and MGC parameter stream, a Mel generalized-log spectrum approximation (MGLSA) filter is utilized. The continuous vocoder is adaptable and straightforward, with only two 1-dimensional parameters for modeling excitation. It can implement its synthesis algorithm in real-time because it is computationally viable. It has the convenience of not requiring a voiced/unvoiced decision, which implies that alignment failures between voiced and unvoiced portions are shunned. But on the other hand, failure in MVF modeling can result in voicing issues. The continuous vocoder is also viable in statistical voice conversion utilizing a feed-forward deep neural network and is comparable to the baseline WORLD vocoder [2].

Furthermore, utilizing a temporal envelope such as Discrete All Pole, Frequency Domain Linear Prediction, Low Pass Filter, and True envelopes, the continuous vocoder accurately simulates the noise component of the excitation source [4]. This feature offers our continuous vocoder the advantage over the most generally used vocoder STRAIGHT [21] because the noise segment of the excitation reference is not precisely modeled in the excitation parameters of STRAIGHT [15].

4.1 Baseline: WORLD vocoder

We chose the WORLD vocoder as a baseline due to its ability to synthesize high-quality voice [29]. The WORLD vocoder employs a source-filter-based speech production model with mixed excitation, alike to the legacy STRAIGHT vocoder [43]. The WORLD vocoder uses DIO, CheapTrick, and PLATINUM algorithms to extract the parameters (F0, spectral envelope, and aperiodic parameter), respectively. The above continuous vocoder uses two one-dimensional parameters to generate the excitation signal, whereas the WORLD system employs a five-dimensional band aperiodicity. Moreover, unlike the continuous vocoder, the WORLD vocoder might carry errors at transition boundaries of voiced and unvoiced parts of speech [4].

5 Methodology

For both systems using the continuous and WORLD vocoders, we performed the methods outlined below. We created an average voice model (AVM), which we then customized for

the target speaker. The open-source Merlin [46] framework was utilized to apply the generic FFNN, and RNN model-based TTS.

5.1 Speech corpus

The research work was carried out using the VCTK-Corpus [42]. This CSTR VCTK Corpus includes voice data from 109 English speakers (62 female) who speak diverse accents. A total of 400 phrases are read by each speaker. The sentences were taken from a newspaper, the rainbow passage, and an elicitation section. The dataset is around 44 hours long in total. All of the recordings were made at a sampling rate of 96 kHz (24 bits) and subsequently down-sampled to 48 kHz. The dataset was created for multi-speaker TTS and speaker-adaptive purposes [10, 31]. We re-sampled the waveforms of the database to 16 kHz. The training experiment employed 85 percent of the data, while the testing and validation experiments used the remaining data.

5.2 Build an average voice model

This strategy is classified as transfer learning, which means that the learnings in one activity can be applied to another similar work. In speech synthesis, an AVM can be trained on a large sample of speakers and then used to adjust to a new speaker with fewer data. The AVM learns how the text and speech features are related. It can change some of the parameters (i.e., weights of the neural networks) for the new speaker. We developed the AVM throughout the range of 9 speakers (six females (p225, p228, p229, p230, p231, and p233) and three males (p226, p227, and p232). We have created the data and directories, labels, acoustic characteristics (MGC, MVF, and lf0) for the continuous and (lf0, mgc, and bap) for the WORLD vocoder. Then we trained the acoustic and duration models. We trained these models using FFNN, LSTM, and GRU.

5.3 Adapt the AVM for the target speakers

To overcome limited data of the target speaker, this model collects universal features from the average voice (multiple speakers) model and is trained independently for the target speaker. For analysing purposes, we did adaptation experiments with four speakers (p234 (female), p236 (female), p237 (male), and p247 (male)). We further trained the average acoustic and duration models for the four target speakers by using FFNN, LSTM, and GRU. Each speaker's adaptation data consists of 400 utterances (about 14 minutes). Finally, we synthesized the sentences using the Continuous vocoders and compared the resulted synthesized speech with the results that we obtained by the baseline vocoder.

5.4 Training topology

Feed-forward neural networks and a couple of versions of the recurrent neural networks (LSTM and GRU) utilized in our study are described in this part. These neural networks are used to predict speech reconstruction acoustic features, and textual and phonetic transcriptions are turned into a sequence of linguistic features as input. Hyperparameters are the same for both systems (Continuous and WORLD vocoders). We engaged a high-performance NVidia Titan X graphics processing unit (GPU) for the networks' training.

1. Feed-Forward Neural Networks

FFNN is the simplest type of neural network that can be used for TTS purposes. We applied the FFNN of six hidden layers with tangent hyperbilic units (1024 neurons, sgd optimizer, 25 epochs, and 265 batch size).

2. Long Short Term Memory networks

   Unlike feed-forward mapping, which is done frame by frame of speech, linguistic attributes are sequence-to-sequence mapped to vocoder parameters in RNN [16].

   At the bottom, we used a four-layer feed-forward neural network. We utilized LSTM on top of the feed-forward layers. The feed-forward layers have 1024 hidden units using a tangent activation function in each layer.

   Meantime, LSTM implementations used 512 units. In addition, Adam optimizer was used. The hyperparameters were set up, including batch size 512, dropout rate 0.1, learning rate 0.001, and training epochs 30.

3. Gated Recurrent Unit

   GRUs are incredibly similar in architecture to LSTM, where both use gates to govern data flow in networks. The differences in their architecture are the GRU does not have a cell state, and LSTM has three gates (input, forget, and output) while GRU has only reset and update gates.

   We used a four-layer feed-forward at the bottom of the neural network design. After that, we utilized GRU on top of the feed-forward layers. The feed-forward layers have 512 hidden units using a tangent activation function in each layer. In addition, GRU executions used 256 units. Moreover, Adam optimizer was used. The hyperparameters were set up, including batch size 256, dropout rate 0.0, learning rate 0.002, and training epochs 30.

The pseudocode of our work is summarized in the following steps:

```
Initialize Continuous vocoder in Merlin
Train the Continuous vocoder using the multiple speakers' dataset
Generate MVF, F0, MGC from the dataset
Adapt the model to the target speaker
Generate MVF, F0, MGC from target speaker dataset
Synthesis speech
Build a baseline vocoder with the same steps (1-6)
Objective and Subjective evaluations for both models
IF the synthesized speech = reasonable quality THEN
        The result is ready!
ELSE
        Optimize the model design
ENDIF
```

6 Results

We used objective and subjective evaluations to test the quality of the synthesized waves of the target speaker. We compared all the results obtained by continuous vocoder with WORLD vocoder. In the objective evaluation, we independently tested our continuous vocoder parameters with speaker adaptation in deep neural networks using Mel-Cepstral Distortion (MCD), F0-correlation (F0-CORR), and spectrogram analysis.

6.1 Objective evaluation

1. MCD(dB):
   We determined the Mel-Cepstral Distortion measure coefficients per the whole model output (1). The MCD is a measure of the difference between two mel cepstra sequences. The notion is that the lower the MCD between synthesized and natural mel cepstral sequences, the closer synthetic speech is to natural speech reproduction. Based on these findings, the efficiency of the continuous vocoder systems against the WORLD baseline scheme could be concluded (Table 1).

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

   x and y are the i^{th} cepstral coefficients of the natural and synthetic speech signals, respectively. Table 1 shows slightly better MCD values of WORLD vocoder with FFNN than continuous vocoder FFNN. In contrast, with LSTM and GRU, the continuous vocoder was slightly better compared to the WORLD vocoder, while being significantly not different. Also, we can observe some speaker dependency, i.e., for two out of the four speakers, the WORLD has lower MCD, while for the other two speakers, the Continuous vocoder has lower MDC values. Thus, we can say that the two vocoders are roughly equivalent in this measure.
2. F0-CORR:
   The similarity between reference and synthesized data is reflected in the correlation (whether they are linearly related or not). Overall, a measurement is taken frame per frame (2). F0-CORR shows promising results for continuous vocoder as the values are closer to 1 and higher than the baseline WORLD vocoder (Table 2). This difference in obtained results might be related to the discontinuous fundamental frequency model of the WORLD vocoder. In terms of F0 correlation, we can observe speaker-dependent differences: in the case of FFNN and LSTM, the continuous vocoder was preferred by three speakers out of four, while with the GRU system, the continuous vocoder was

Table 1 MCD errors on the dev/test sets for WORLD & continuous vocoders

Spkr	WORLD vocoder FFNN	WORLD vocoder LSTM	WORLD vocoder GRU
P234	5.232 / 5.246	5.466 / 5.460	5.474 /5.457
P236	5.601 / 5.297	5.819 / 5.578	5.854 / 5.702
P237	5.139 / 5.030	5.487 / 5.444	5.298 / 5.186
P247	5.360 / 5.342	5.517 / 5.530	5.566 / 5.577
Average	5.333 / 5.228	5.572 / 5.503	5.548/5.480
	Cont. vocoder	Cont. vocoder	Cont. vocoder
P234	5.440 / 5.434	5.541 / 5.492	5.480 / 5.378
P236	5.795 / 5.547	5.817 / 5.554	5.805 / 5.518
P237	5.370 / 5.240	5.283 / 5.194	5.303 / 5.203
P247	5.502 / 5.483	5.537 / 5.512	5.446 / 5.491
Average	5.526 / 5.426	5.544 / 5.438	5.508 / 5.397

Table 2 F0-CORR on the dev/test sets for WORLD & continuous vocoders

Spkr	WORLD vocoder FFNN	WORLD vocoder LSTM	WORLD vocoder GRU
P234	0.484 / 0.463	0.460 / 0.408	0.428 / 0.387
P236	0.580 / 0.565	0.548 / 0.552	0.486 / 0.489
P237	0.597 / 0.551	0.514 / 0.483	0.515 / 0.540
P247	0.588 / 0.654	0.538 / 0.622	0.547 / 0.628
Average	0.562 / 0.558	0.515 / 0.516	0.494 / 0.511
	Cont. vocoder	Cont. vocoder	Cont. vocoder
P234	0.730 / 0.755	0.737 / 0.749	0.751 / 0.759
P236	0.480 / 0.588	0.508 / 0.572	0.505 / 0.507
P237	0.760 / 0.721	0.739 / 0.709	0.718 / 0.710
P247	0.766 / 0.682	0.769 / 0.696	0.773 / 0.693
Average	0.684 / 0.686	0.688 / 0.681	0.686 / 0.667

preferred by all four speakers. Thus, we can conclude that the continuous vocoder was ranked as being better than the WORLD baseline in terms of F0 modeling. This can be explained by the way they model F0 (see Section 3).

$$F0 - CORR = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{2}$$

Where x_i is the natural speech frame, y_i is the synthesized speech frame, $\overline{x}$ is the mean of the natural speech frame, and $\overline{y}$ is the mean of the synthesized speech frame.

3. Demonstration sample (spectrogram):

   Figure 3 shows the spectrograms of synthesized sounds by both vocoders (WORLD and continuous). We present six spectrogram plots of three speakers of two females (p237, p236) and a male (p247) reading the text "People look, but no one ever finds it.". As can be seen in the spectrograms in the right column, the continuous vocoder distinguishes between voiced and unvoiced speech frequencies based on the MVF parameter (i.e., spectral content below MVF is voiced, while spectral range above MVF is unvoiced). The spectrograms in the left column do not show such a strong separation curve in terms of frequency, because the WORLD vocoder handles voicing via band aperiodicities. Thus, the main difference between the baseline WORLD and the proposed continuous vocoder is in the way they model the excitation component of speech.

6.2 Subjective evaluation

We ran an online MUSHRA-like test to analyze TTS variants [19]. This investigation compared the synthesized phrases generated by the baseline (WORLD) and the proposed (Continuous) vocoders. Natural versions of the sentences were included as references. Besides, we created a lower anchor variant, a spectrally distorted version of the natural

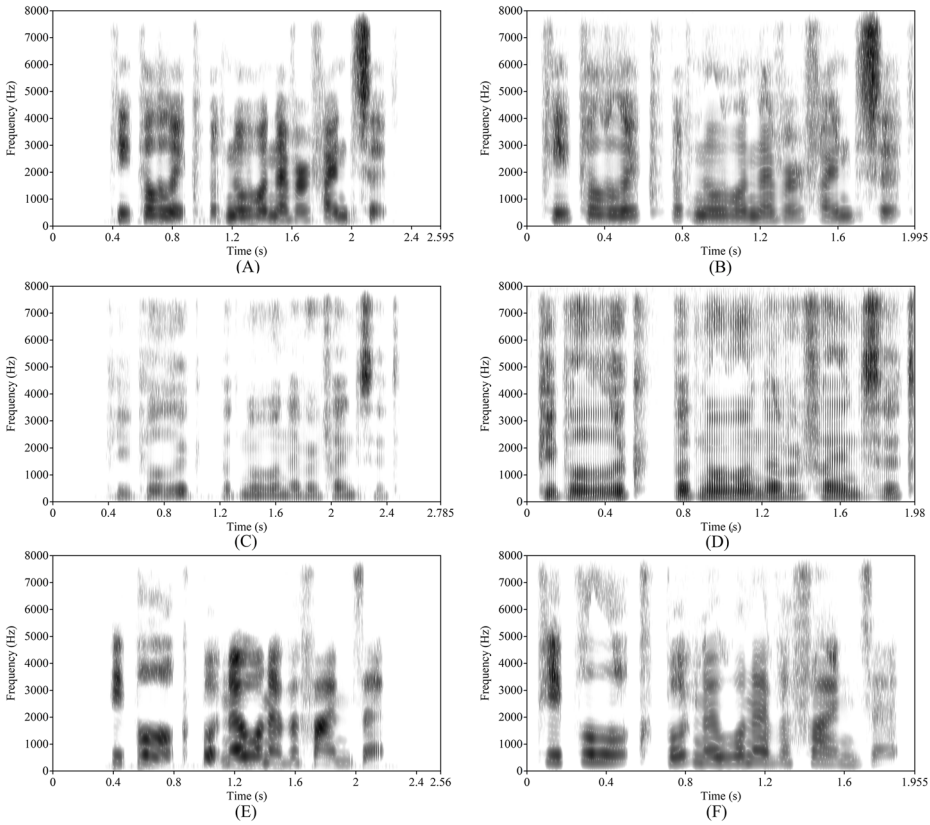

Fig. 3 The spectrogram plot of six synthesized sounds (a) a synthesized female voice (p237) using WORLD vocoder with FFNN, (b) a synthesized female voice (p237) using continuous vocoder with FFNN, (c) a synthesized male voice (p247) using WORLD vocoder with LSTM, (d) a synthesized male voice (p247) using continuous vocoder with LSTM, (e) a synthesized female voice (p236) using WORLD vocoder with GRU, and (f) a synthesized female voice (p236) using continuous vocoder with GRU

sentence. Subjects were requested to score the naturalness of each stimulus on a scale of 0 (very unnatural) to 100 (fully natural or realistic). We generated 19 sentences from the VCTK-Corpus using the four target speakers, of which 12 were selected for the test. The samples were displayed in a random order (different for each participant). The MUSHRA test comprised 96 utterances (8 systems x 4 speakers x 12 sentences).

Audiences were provided an example to listen to before the test to adjust the loudness. Twenty one non-native English speakers (17 silent situations, 4 in a noisy environment; one female, 20 males; 26-41 years old) scored each sentence. The test took an average of 18 minutes to achieve. Figure 4 displays the average naturalness scores for the approaches that were tried. Original utterances received 87 out of 100 scores from the listeners. The synthesized speech using WORLD vocoder with FFNN obtained higher scores (52 out of 100) than our model using the same network topology (39). On the other hand, both systems with GRU received almost the same votes (57 and 50, respectively). The continuous vocoder showed an almost equivalent performance (53 out of 100) to the WORLD vocoder (59) by using LSTM.

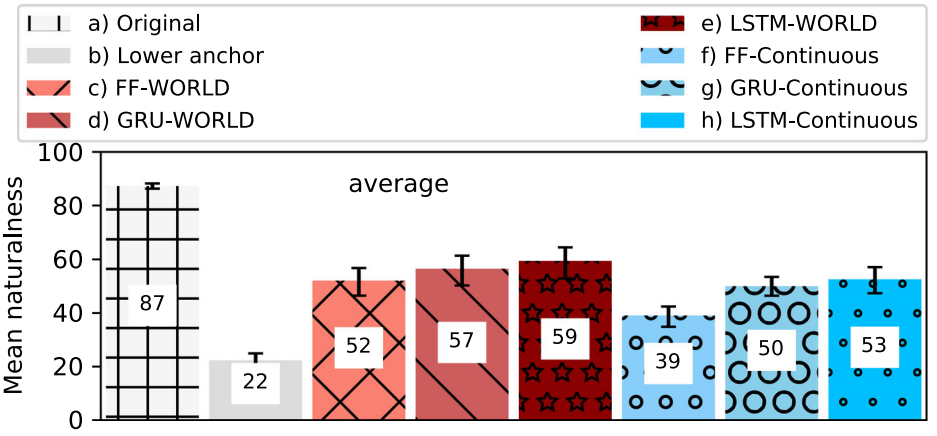

Fig. 4 MUSHRA scores for the naturalness question – average results. Higher value means better overall quality. Errorbars show the bootstrapped 95% confidence intervals

Figure 5 shows the MUSHRA test results speaker by speaker. For all speakers, the WORLD vocoder with FFNN was preferred over the continuous vocoder. GRU presents a different scenario. Both systems are almost equivalent in the higher f234, f236, and m247. With the speaker m237, the WORLD vocoder obtained more evaluation scores than our system. Moreover, WORLD vocoder obtained more scores with a synthesized speech by the three speakers (f234, f237, and m247) than continuous vocoder using LSTM except for one speaker (f236), which shows equal results. These disparities are statistically significant (Mann-Whitney-Wilcoxon ranksum test, with a 95% confidence level). Despite the fact that the naturalness scores of the Continuous vocoder did not match those of the WORLD vocoder, we may infer that the speaker adaptation experiment with the suggested vocoder was a success and we obtained reasonable results.

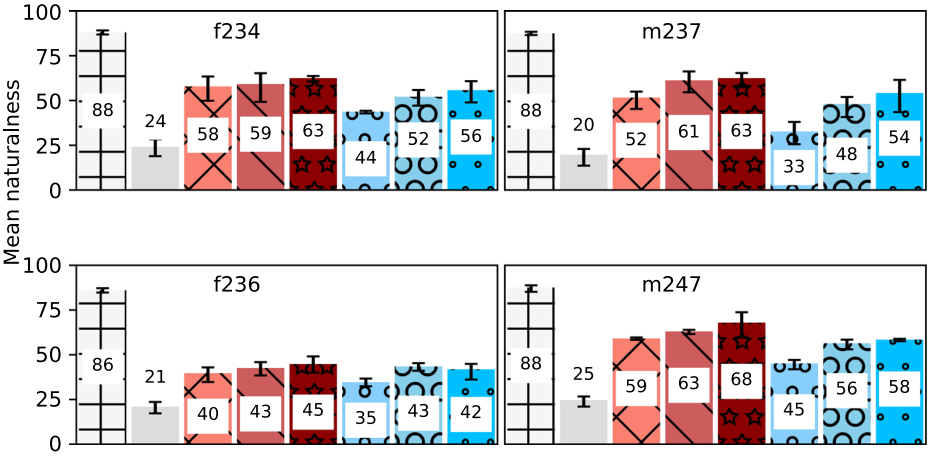

Fig. 5 MUSHRA scores for the naturalness question – speaker by speaker

7 Conclusions and future work

Recent neural vocoders, while being capable of synthesizing highly natural speech, often also fall short of the criteria of real-time synthesis. Classical vocoder-based SPSS can be effective in applications demanding modest computer complexity (e.g., a chatbot in a high-load environment like a government website, or a low-resource smartphone). This research aimed to present a new strategy for improving the accuracy of our prior continuous vocoder with speaker adaptation. We employed the continuous vocoder, in which the excitation is characterized with two one-dimensional parameters: continuous F0 and MVF. During our evaluation experiments, we prove that using data from multiple speakers it is achievable to train an average voice model and that speaker adaptation is possible with just 400 utterances (roughly 14 minutes). The performance strengths and shortcomings of each of the proposed approaches (FFNN, GRU, and LSTM) for four different speakers were analyzed using a number of metrics. The objective testing revealed that our vocoder is capable of synthesizing voices using RNN topologies roughly equivalent to the baseline WORLD vocoder. Furthermore, the MUSHRA test results proved the efficacy of our speaker adaptation strategy, being only slightly worse than the WORLD vocoder in the case of GRU and LSTM. The WORLD vocoder obtained significantly higher scores than our system in the case of FFNN. Text-to-speech synthesis with a low-complexity vocoder and speaker adaptation can help more natural human-computer interaction.

In future research, we will explore the speaker adaptation with the Continuous vocoder using a hybrid model (e.g., RNN, CNN, gcForest, etc.). These models deliver more effectively than the individual deep neural networks. Moreover, Transformer neural networks, a state-of-the-art technique in natural language processing (NLP), with our model is our target for the next search step. In addition, building a multi-model speaker adaptation system of the Continuous vocoder with state-of-the-art sequence-to-sequence models such as Tacotron2, FastPitch, etc., more advanced solutions would be considered.

Acknowledgements The research was partially sponsored by the APH-ALARM project (contract 2019-2.1.2-NEMZ-2020-00012), funded by the European Commission and the National Research, Development and Innovation Office of Hungary and supported by the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory. The research reported in this publication, carried out by the Department of Telecommunications and Media Informatics Budapest University of Technology and Economic and IdomSoft Ltd., was supported by the Ministry of Innovation and Technology and the National Research, Development and Innovation Office within the framework of the National Laboratory of Infocommunication and Information Technology. Tamás Gábor Csapó's research was supported by the Bolyai János Research Fellowship of the Hungarian Academy of Sciences and by the ÚNKP-21-5 (identifier: ÚNKP-21-5-BME-352) New National Excellence Program of the Ministry for Innovation and Technology from the source of the National, Research, Development and Innovation Fund. 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.

Declarations

Conflict of Interests The authors have no relevant financial interests in the manuscript and no other potential conflicts of interest.

References

1. 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: Interspeech, pp 434–438
2. Al-Radhi MS, Csapó TG, Németh G (2019) Continuous vocoder applied in deep neural network based voice conversion. Multimed Tools Appl 78(23):33549–33572
3. Al-Radhi MS, Abdo O, Csapó TG, Abdou S, Németh G, Fashal M (2020) A continuous vocoder for statistical parametric speech synthesis and its evaluation using an audio-visual phonetically annotated arabic corpus. Comput Speech Lang 60:101025
4. Al-Radhi MS, Csapó TG, Németh G (2021) Noise and acoustic modeling with waveform generator in text-to-speech and neutral speech conversion. Multimed Tools Appl 80(2):1969–1994
5. Atkar G, Jayaraju P (2021) Speech synthesis using generative adversarial network for improving readability of Hindi words to recuperate from dyslexia. Neural Comput Applic 33(15):9353–9362
6. Babacan O, Drugman T, Raitio T, Erro D, Dutoit T (2014) Parametric representation for singing voice synthesis: a comparative evaluation. In: ICASSP. IEEE, pp 2564–2568
7. Bollepalli B, Juvela L, Airaksinen M, Valentini-Botinhao C, Alku P (2019) Normal-to-lombard adaptation of speech synthesis using long short-term memory recurrent neural networks. Speech Comm 110:64–75
8. Casanova E, Shulby C, Gölge E, Müller NM, de Oliveira FS, Candido AJ Jr, da Silva Soares A, Aluisio SM, Ponti MA (2021) SC-Glow TTS: an efficient zero-shot multi-speaker text-to-speech model. In: Proceedings of the interspeech 2021, pp 3645–3649
9. Chen Y, Assael Y, Shillingford B, Budden D, Reed S, Zen H, Wang Q, Cobo LC, Trask A, Laurie B (2019) Sample efficient adaptive text-to-speech. In: International conference on learning representations (ICLR)
10. Chen M, Tan X, Ren Y, Xu J, Sun H, Zhao S, Qin T, Liu Tie-Yan (2020) Multispeech: multi-speaker text to speech with transformer. Interspeech:4024–4028
11. Chen B, Du C, Yu K (2022) Neural fusion for voice cloning. IEEE/ACM Trans Audio Comput Speech Lang 30:1993–2001
12. Choi S, Han S, Kim D, Ha S (2020) Attentron: few-shot text-to-speech utilizing attention-based variable-length embedding. In: Proceedings of the interspeech 2020, pp 2007–2011
13. Csapó TG, Németh G, Cernak M, Garner PN (2016) Modeling unvoiced sounds in statistical parametric speech synthesis with a continuous vocoder. In: EUSIPCO. IEEE, pp 1338–1342
14. Dai D, Chen Y, Chen L, Tu M, Liu L, Xia R, Tian Q, Wang Y, Wang Y (2022) Cloning one's voice using very limited data in the wild. In: ICASSP, pp 8322–8326
15. Degottex G, Erro D (2014) A uniform phase representation for the harmonic model in speech synthesis applications. EURASIP Journal on Audio, Speech, and Music Processing 2014(1):1–16
16. Fan Y, Qian Y, Xie F-L, Soong FK (2014) Tts synthesis with bidirectional lstm based recurrent neural networks. In: Interspeech
17. Hinterleitner F (2017) Speech synthesis. In: Quality of synthetic speech. Springer, pp 5–18
18. Hu Q, Richmond K, Yamagishi J, Latorre J (2013) An experimental comparison of multiple vocoder types. In: SSW8
19. ITU-R recommendation BS.1534: method for the subjective assessment of intermediate audio quality, 2001
20. Jia Y, Zhang Y, Weiss RJ, Wang Q, Shen J, Ren F, Chen Z, Nguyen P, Pang R, Moreno IL et al (2018) Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In: Conference on neural information processing systems (NIPS). arXiv:1806.04558
21. Kawahara H, Masuda-Katsuse I, De Cheveigne A (1999) Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: possible role of a repetitive structure in sounds. Speech Comm 27(3-4):187–207
22. Latif S, Qadir J, Qayyum A, Usama M, Younis S (2020) Speech technology for healthcare opportunities, challenges, and state of the art. IEEE Rev Biomed Eng 14:342–356
23. Latif S, Qadir J, Qayyum A, Usama M, Younis S (2021) Speech technology for healthcare opportunities, challenges, and state of the art. IEEE Rev Biomed Eng 14:342–356
24. Lee J-H, Lee S-H, Kim J-H, Lee S-W (2022) PVAE-TTS adaptive text-to-speech via progressive style adaptation. In: ICASSP, pp 6312–6316

25. Li X, Ma D, Yin B (2021) Advance research in agricultural text-to-speech: the word segmentation of analytic language and the deep learning-based end-to-end system. Comput Electron Agric 180:105908
26. Luong H-T, Yamagishi J (2018) Scaling and bias codes for modeling speaker-adaptive dnn-based speech synthesis systems. In: SLT. IEEE, pp 610–617
27. Luong H-T, Yamagishi J (2020) Nautilus: a versatile voice cloning system. IEEE/ACM Transactions on Audio Speech, and Language Processing 28:2967–2981
28. Mandeel AR, Mohammed Salah AR, Csapó TG (2021) Speaker adaptation with continuous vocoder-based dnn-tts. SPECOM lecture notes in computer science, vol 12997. Springer, Cham
29. Morise M, Yokomori F, Ozawa K (2016) World: a vocoder-based high-quality speech synthesis system for real-time applications. IEICE Trans Inf Syst 99(7):1877–1884
30. Ning Y, He S, Wu Z, Xing C, Zhang L-J (2019) A review of deep learning based speech synthesis. Appl Sci 9(19):4050
31. Ping W, Peng K, Gibiansky A, Arik SO, Kannan A, Narang S, Raiman J, Miller J (2018) Deep voice 3: scaling text-to-speech with convolutional sequence learning. In: International conference on learning representations (ICLR)
32. Prenger R, Valle R, Catanzaro B (2019) Waveglow: a flow-based generative network for speech synthesis. In: ICASSP. IEEE, pp 3617–3621
33. Quatieri TF (2006) Discrete-time speech signal processing: principles and practice. Pearson Education, India
34. Rao A, Ghosh PK (2020) Sfnet: a computationally efficient source filter model based neural speech synthesis. IEEE Signal Process Lett 27:1170–1174
35. Schnell B, Garner PN (2022) Investigating a neural all pass warp in modern TTS applications. Speech Comm 138:26–37
36. Senior A, Lopez-Moreno I (2014) Improving dnn speaker independence with i-vector inputs. In: ICASSP. IEEE, pp 225–229
37. Shiga Y, Ni J, Tachibana K, Okamoto T (2020) Text-to-speech synthesis. In: Speech-to-speech translation. Springer, pp 39–52
38. Silvestri R, Holmes A, Rahemtulla R (2021) The interaction of cognitive profiles and text-to-speech software on reading comprehension of adolescents with reading challenges. J Spec Educ Technol 0(0):01626434211033577
39. Takaki S, Kim S, Yamagishi J (2016) Speaker adaptation of various components in deep neural network based speech synthesis. In: SSW, pp 153–159
40. Tejedor-García C, Escudero-Mancebo D, Cámara-Arenas E, González-Ferreras C, Cardeñoso-Payo V (2020) Assessing pronunciation improvement in students of english using a controlled computer-assisted pronunciation tool. IEEE Trans Learn Technol 13(2):269–282
41. van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K (2016) Wavenet: a generative model for raw audio. CoRR, vol.abs/1609.0
42. Veaux C, Yamagishi J, MacDonald K et al (2017) Cstr vctk corpus: english multi-speaker corpus for cstr voice cloning toolkit. University of Edinburgh. The Centre for Speech Technology Research (CSTR)
43. Wang X, Lorenzo-Trueba J, Takaki S, Juvela L, Yamagishi J (2018) A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis. In: ICASSP. IEEE, pp 4804–4808,
44. Wu Z, Swietojanski P, Veaux C, Renals S, King S (2015) A study of speaker adaptation for dnn-based speech synthesis. In: Interspeech
45. 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: ICASSP. IEEE, pp 4460–4464
46. Wu Z, Watts O, King S (2016) Merlin: an open source neural network speech synthesis system. In: SSW, pp 202–207
47. Xie X, Liu X, Lee T, Wang L (2021) Bayesian learning for deep neural network adaptation. IEEE/ACM Transactions on Audio, Speech, and Language Processing
48. Xue S, Abdel-Hamid O, Jiang H, Dai L, Liu Q (2014) Fast adaptation of deep neural network based on discriminant codes for speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing 22(12):1713–1725
49. Yang S, Xie L, Chen X, Lou X, Zhu X, Huang D, Li H (2017) Statistical parametric speech synthesis using generative adversarial networks under a multi-task learning framework. In: 2017 IEEE automatic speech recognition and understanding workshop (ASRU), pp 685–691
50. Zhang H, Lin Y (2022) Improve few-shot voice cloning using multi-modal learning. In: ICASSP, pp 8317–8321