Continuous vocoder in feed-forward deep neural network based speech synthesis

Mohammed Salah Al-Radhi, Tamás Gábor Csapó, and Géza Németh

Abstract—Recently in statistical parametric speech synthesis, we proposed a vocoder using continuous F0 in combination with Maximum Voiced Frequency (MVF), which was successfully used with hidden Markov model (HMM) based text-to-speech (TTS). However, HMMs often generate over-smoothed and muffled synthesized speech. From this point, we propose here to use the modified version of our continuous vocoder with deep neural networks (DNNs) for further improving its quality. Evaluations between DNN-TTS using Continuous and WORLD vocoders are also presented. Experimental results from objective and subjective tests have shown that the DNN-TTS have higher naturalness than HMM-TTS, and the proposed framework provides quality similar to the WORLD vocoder, while being simpler in terms of the number of excitation parameters and models better the voiced/unvoiced speech regions than the WORLD vocoder.

Index Terms—continuous vocoder, feed-forward neural networks, speech synthesis, V/UV.

I. BACKGROUND

In the last decade, a large number of vocoders have been proposed (for a comparison, see [1]) for statistical parametric speech synthesis. The direct antecedent of the current work had been carried out by [2] which proposed a computationally feasible residual-based vocoder. During the analysis phase, fundamental frequency (F0) is calculated on the input waveforms of a simple continuous pitch tracker [3]. In regions of creaky voice and in case of unvoiced sounds or silences, this pitch tracker interpolates F0 based on a linear dynamic system and Kalman smoothing. After this step, Maximum Voiced Frequency (MVF) is calculated from the speech signal [4], resulting in the MVF parameter. In the next step, 24-order Mel-Generalized Cepstral analysis (MGC) [5] is performed on the speech signal with alpha=0.42 and gamma=1/3. In all steps, 5 ms frame shift is used. The results are the F0cont, MVF and the MGC parameter streams. Finally, the baseline system performs Principal Component Analysis (PCA) on the pitch synchronous residuals.

During the synthesis phase of the baseline system, voiced excitation is composed of PCA residuals overlap-added pitch synchronously, depending on the continuous F0. After that, this 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. Voiced and unvoiced excitation is added together. Finally, an MGLSA filter is used to synthesize speech from the excitation and the MGC parameter stream [6].

The primary goal of this paper is to model the improved version of the continuous vocoder parameters (F0, MVF, and MGC) with feed-forward DNN based speech synthesis. Besides, noise components in voiced sounds are parameterized and modeled to meet the requirements of high sound quality. Finally, an experimental comparison between our Continuous and the WORLD vocoders is presented.

II. PROPOSED METHODS

A. DNN based speech synthesis

Figure 1 conceptually illustrates the main components of the Continuous vocoder when applied in DNN-based training. Textual and phonetic parameters are first converted to a sequence of linguistic features as input, and neural networks are employed to predict acoustic features as output for synthesizing speech. The DNN applied here is a feed-forward multilayer perceptron architecture. The input is used to predict the output with six layers of hidden units, each of which performs a non-linear function of the previous layer’s representation, and a linear activation function was used at the output layer. We applied a hyperbolic tangent activation function whose outputs lie in the range (-1 to 1) which 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 of 0.3, and 10 epochs later, the momentum was increased to 0.9 and then the learning rate was halved regularly. The DNN used in this research was implemented in the Merlin open source neural network speech toolkit [7].

B. Baseline vocoder

Earlier vocoders have been developed to synthesize high-quality speech based on DNN training models. In this work, WORLD vocoder was chosen as a baseline for comparison with our optimized vocoder.

Mohammed Salah Al-Radhi is with the Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary (e-mail: malradhi@tmit.bme.hu).

Tamás Gábor Csapó is with the Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics. He is also with the MTA-ELTE Lendület Lingual Articulation Research Group, Budapest, Hungary (e-mail: csapo@tmit.bme.hu).

Géza Németh is with the Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary (e-mail: nemeth@tmit.bme.hu).
The improved version of the WORLD vocoder was proposed in [8]. Similarly to other approaches, the WORLD vocoder is based on source-filter separation, i.e. models separately the excitation (with F0 and aperiodicity) and the spectral envelope. At the beginning, WORLD estimates the F0 contour using the DIO (Distributed Inline-filter Operation) algorithm [9]. Then, the excitation signal is estimated with the D4C (Definitive Decomposition Derived Dirt-Cheap) algorithm [10], and used as a band aperiodicity of speech signals.

C. Improved version of continuous vocoder

Degottex et al. argue that the noise component is not accurately modeled in modern vocoders (even in the widely used STRAIGHT vocoder) [11]. In the standard Continuous vocoder [2], there is a lack of voiced components in higher frequencies. However, it was shown that in natural speech, the high-frequency noise component is time-aligned with the pitch periods [12]. For this reason, we applied several time envelopes to shape the high-frequency noise excitation component (above MVF) [13]. From the several envelopes investigated, the True envelope was found to be the best. Therefore, this will be used in the current study. Moreover, we have also proposed that the True envelope with weighting factor will bring us a unique time envelope which makes the convergence more close to natural speech. In practice, the weight factor which was found to be the most successful is 10.

In our recent study [2] [13], a simple spectral model represented by 24-order mel-generalized cepstral coefficients was used [5]. However, more advanced spectral estimation methods might increase the quality of synthesized speech. In [14], an accurate and temporally stable spectral envelope estimation is called CheapTrick has been proposed. CheapTrick consists of three steps: F0-adaptive Hanning window, smoothing of the power spectrum, and spectral recovery in the quefrency domain. Hence, Cheaptrick algorithm based on the 60-order MGC representation will be used as a unified approach to speech spectral estimation in a modified version of continuous vocoder.

Table 1 compares the parameters of the vocoders under study. It can be seen that the continuous vocoder uses only two one-dimensional parameters for modeling the excitation, whereas the WORLD vocoder is applying a five-dimensional band aperiodicity. Also, for the DNN training with the WORLD vocoder, it is necessary to interpolate F0 and add a new voiced/unvoiced binary feature.

<table>
<thead>
<tr>
<th>Vocoder</th>
<th>Parameters per frame</th>
<th>Excitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>F0: 1 + MVF: 1</td>
<td>Mixed</td>
</tr>
<tr>
<td></td>
<td>+ MGC: 60</td>
<td></td>
</tr>
<tr>
<td>WORLD</td>
<td>F0: 1 + Band aperiodicity: 5</td>
<td>Mixed</td>
</tr>
<tr>
<td></td>
<td>+ MGC: 60</td>
<td></td>
</tr>
</tbody>
</table>

### III. EVALUATION AND DISCUSSION

In order to achieve our goals and to verify the effectiveness of the proposed methods, objective and subjective evaluations were carried out. We conducted two kinds of experimental evaluations. In the first evaluation, we compared our proposed Continuous vocoder with the WORLD vocoder in terms of the F0 modeling capability. In the second evaluation, we tested them using a subjective listening experiment.

A. Data

To measure the performance of the obtained model, the US English female (SLT) speaker was chosen for the experiment from the CMU-ARCTIC database [15], which consists of 1132 sentences. 90% of the sentences were used for training and the rest was used for testing.

B. F0 modeling capability

To evaluate the performance of the WORLD and Continuous vocoders, the F0 modeling capability and the V/UV transitions were tested the following way. Although the WORLD vocoder can achieve a high quality
when applied in speech synthesis, it is worth noting here that the WORLD vocoder (which is using the DIO pitch tracking algorithm and results in a discontinuous F0 track) can make V/UV decision errors (i.e. setting voiced that should be unvoiced, or vice versa) and also sometimes contains errors at boundaries (at the V/UV or UV/V transitions). This is not the case with the Continuous vocoder, which is using a continuous pitch detection algorithm. In the latter, the voicing feature is modeled by the continuous Maximum Voiced Frequency parameter; therefore, V/UV errors do not occur. It can be seen in Figure 3 (showing the F0 contour of a short segment of synthesized speech) that the Continuous vocoder interpolates the F0 contour even in unvoiced regions of speech. According to the final results with the DNN, the V/UV error was 5.35% for the WORLD vocoder in case of the SLT speaker. In informal listening tests we also observed that the baseline vocoder often synthesizes speech with clicks which are the result of false V/UV decisions.

![F0 trajectories](image)

Fig. 2. F0 trajectories for the synthesized speech signal using the DIO algorithm (red), and Continuous algorithm (blue) for WORLD and Continuous vocoders respectively. (sentence: “Author of the danger trail, Philip Steels, etc.”, from speaker SLT).

C. RMS - Log Spectral Distance

To verify the effectiveness of the proposed vocoder by CheapTrick algorithm in the direction of refining baseline vocoder spectral envelope [2] [13], root mean square (RMS) log spectral distance (LSD) evaluation was carried out. RMS-LSD is a distance measure and can be defined here by

$$LSD = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \text{mean} \left[ \log P(f_i) - \log \hat{P}(f_i) \right]^2}$$

where $P(f)$ and $\hat{P}(f)$ are spectral power magnitudes of the nature and synthesis speech respectively, defined at N frequency points.

For a perfect synthesized speech, the ideal value of LSD is zero, which indicates matching frequency content. The results from our experiment confirm that LSD is getting lower by using CheapTrick spectral algorithm than the simple spectral algorithm used in the baseline vocoder. This point is well illustrated in Figure 3 by three spectrograms of frequency versus time. In the middle spectrum, we have detected a spectrum with LSD equal to 1.5, while the bottom spectrum has a better LSD equal to 0.92 that is almost the same in comparison to the top speech spectrum (natural sound). Thus, we can say that our proposed scheme introduces small distortion to the sound quality and satisfies correct spectral criterion.

![Spectrograms](image)

Fig. 3. Comparison of the speech spectrums synthesized by proposed continuous vocoder: Natural speech signal (top), synthesized speech based on a simple MGC algorithm (middle), and synthesized speech based on CheapTrick algorithm (bottom). The sentence is “He made sure that the magazine was loaded, and resumed his paddling.” from speaker SLT).

D. Subjective listening test

In order to evaluate the differences in DNN-TTS synthesized samples using the above vocoders, we conducted a web-based MUSHRA (MUlti-Stimulus test with Hidden Reference and Anchor) listening test [16]. We compared natural sentences with the synthesized sentences from the baseline, proposed and a benchmark system. The benchmark was a DNN-TTS applied with a simple pulse-noise excitation vocoder. Also, we added samples from an earlier HMM-TTS system which was using the Continuous vocoder [2]. 15 sentences were selected from the SLT speaker which were not included in the training. Altogether, 90 utterances were included in the test (6 types x 15 sentences). In the test, the listeners had to rate the naturalness of each stimulus relative to the reference (which was the natural sentence), from 0 (highly unnatural) to 100 (highly natural). The utterances were presented in a randomized order (different for each participant). The listening test samples can be found online:http://smartlab.tmit.bme.hu/dogs2017_vocoder_dnn

Nine participants (7 males, 2 females) with a mean age of 35 years, mostly with engineering background were asked to conduct the online listening test. On average, the test took 20 minutes to fill. The results of the listening test are presented in Figure 4. According to the results, all DNN
systems outperformed the HMM system. From the three vocoders applied in DNN-TTS, the WORLD vocoder reached the highest scores, which is significantly different from the other systems (Mann-Whitney-Wilcoxon ranksum test, p<0.05). The difference between the two types of the Continuous vocoder is not significant. From this test, we can conclude that 1) the DNN-TTS with the Continuous vocoder is more natural than the HMM-TTS, 2) the WORLD vocoder was rated better than the Continuous vocoder.

IV. CONCLUSIONS

The goal of the work reported in this paper was to apply a Continuous vocoder in deep neural network based speech synthesis. The experiments were successful and we were able to add the continuous features (F0, Maximum Voiced Frequency, and CheapTrick Mel-Generalized Cepstrum) to the training of the DNNs. One English speaker was tested with a baseline (WORLD) vocoder and two versions of the Continuous vocoder. The motivation for using a Continuous vocoder arises from our observation that the WORLD vocoder has often V/UV errors and boundary errors due to the DIO F0 estimation algorithm. In a subjective listening test, we found that the DNN-TTS using the Continuous vocoder was rated better than an earlier HMM-TTS system. However, the proposed vocoders were not found to be better than the baseline vocoder when applied in DNN-based speech synthesis. The reason for this might be that the modeling of the high-frequency excitation components is still not optimal in the proposed vocoder. We plan to address this in the future. As the Continuous vocoder has few parameters and is computationally feasible, it is suitable for real-time operation once the DNN-based parameter generation will be fast enough for this.

For future work, the authors plan to investigate the effectiveness of applying the performance of a mixture density recurrent network by combing with bi-directional long-short memory (Bi-LSTM) based TTS to further improve and refine continuous parameters.

V. ACKNOWLEDGEMENTS

The research was partly supported by the VUK (AAL-2014-1-183) and the EUREKA / DANSPLAT projects. We would like to thank the subjects for participating in the listening test.

REFERENCES