conTTS: Text-to-Speech Application using a Continuous Vocoder

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

[1]Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics, Budapest, Hungary
[2]MTA-ELTE Lendület Lingual Articulation Research Group, Budapest, Hungary
{malradhi,csapot,nemeth}@tmit.bme.hu

1. Introduction

A Text-to-Speech (TTS) system generates speech from the corresponding text and used in various educational, telecommunication and multimedia applications. Recently, with the rapid development of artificial intelligence technologies, end-to-end generative TTS models (such as WaveNet, Tacotron, char2wav) are proposed to predict speech parameters in a unified way, which makes the quality and naturalness of synthesized speech greatly improved. However, problems can still occur, e.g. wrong stress patterns, unreasonable breaks, and require a large amount of speech data from one speaker to obtain good quality. Therefore, parametric vocoding is central to the success of state-of-the-art TTS systems.

2. Proposed Methodology

Here we describe the conTTS system, a novel TTS synthesizer based on a new open-source continuous vocoder which is simple to get started with, but also offers advanced features and achieves high-quality [1]. The conTTS design and performance are interactive and do not require any special skills while being faster than neural vocoders used in state-of-the-art TTS systems. The developed system supports a preliminary evaluation on English speakers that could assist people with speech disorders. We map linguistic features containing phone, syllable and duration information to acoustic mel cepstral features (MGC) [2], pitch (F0) [3], and maximum voiced frequency (MVF) [4] using DNNs. This toolkit allows users to enter their sentences as text and automatically obtain the synthesized speech output. It is available under an open source license[1], and the overall system architecture is illustrated in Figure 1.

The acoustic model was built using the recent Merlin toolkit. In this work, we have used the feed-forward DNN and RNNs (LSTM, BLSTM, and GRU) based statistical parametric speech synthesis. Additionally, our model is simpler to train and yields sharper acoustic models. The Festival toolkit is used to extract the linguistic features, which encode information regarding the phone identity, quinphone context, and syllable stress features of adjacent syllables. Our training corpus consists of around 1150 utterances[2] per each speaker, recorded by US English male and female speakers, both experienced voice talents. We have divided the total dataset into train, validation and test sets. The model is evaluated on test data and observed a high-quality training accuracy.

3. Discussion and Evaluation

In a given English text sentence, users can select one of two voice patterns (either male or female) from the current set to build their custom voice model. Users can also specify the neural network topology to be trained as well as the number of hidden layers. Synthesized speech samples generated by conTTS can be found online[3]. Figure 2 is an example of the spectral envelope extracted from a speech frame, whereas Figure 3 shows the spectrograms of a synthesized speech with MVF.

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, the WORLD vocoder applies five-dimensional band aperiodicity, while STRAIGHT computes high-dimensional parameters which makes the statistical modelling approach progressively complex and computationally intensive.

[1] https://github.com/malradhi/merlin
[2] http://www.festvox.org/cmu_arctic/cmuarctic.data
[3] https://malradhi.github.io/conTTS/

The findings also point out that the continuous vocoder has few parameters compared to the WORLD and STRAIGHT vocoders, and it is computationally feasible; therefore, it is suitable for real-time operation.

4. Conclusions and Future Work

In this work, we have presented the conTTS framework. It is a speech analysis and synthesis system, that is lightweight and easy to use. We demonstrate that our method can provide comparatively simple waveform TTS synthesis. Future work might be focused on voice conversion ensuring compatibility with a non-parallel dataset.

5. Acknowledgements

The research was partly supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825619 (AI4EU), and by the National Research Development and Innovation Office of Hungary (FK 124584 and PD 127915). The Titan X GPU used was donated by NVIDIA Corporation.

6. References

[1] M.S. Al-Radhi, et al, A continuous vocoder for statistical parametric speech synthesis and its evaluation using an audio-visual phonetically annotated Arabic corpus, *Computer Speech and Language*, 60, pp. 1-15, 2020.

[2] M. Morise, CheapTrick, a spectral envelope estimator for high-quality speech synthesis, *Speech Communication*, vol. 67, pp. 1-7, 2015.

[3] P.N. Garner, M. Cernak, and P. Motlicek, A simple continuous pitch estimation algorithm, *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 102-105, 2013.

[4] T. Drugman and Y. Stylianou, Maximum Voiced Frequency Estimation: Exploiting Amplitude and Phase Spectra, *IEEE Signal Processing Letters*, vol. 21, no. 10, pp. 1230–1234, 2014.

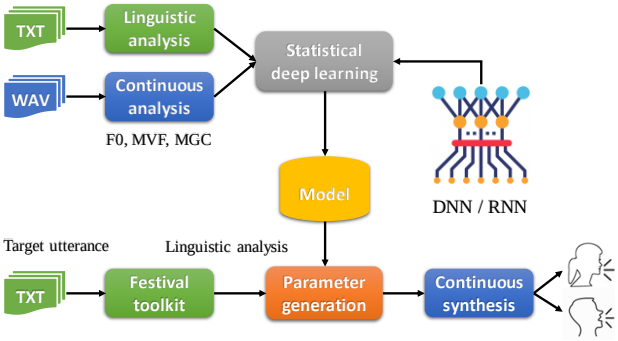

Figure 1: *Schematic overview of the conTTS system flow.*

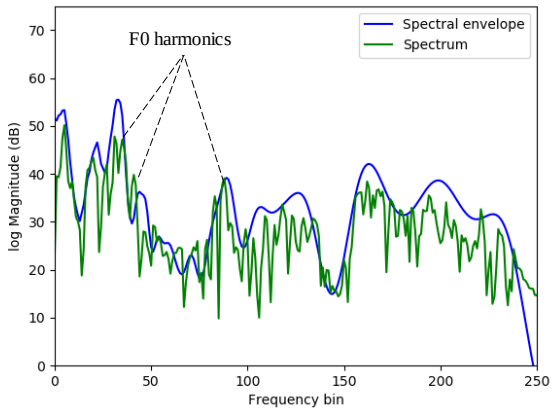

Figure 2: *Within a speech frame, the continuous vocoder measures the F0 and the MGC representation of the spectral envelope (blue).*

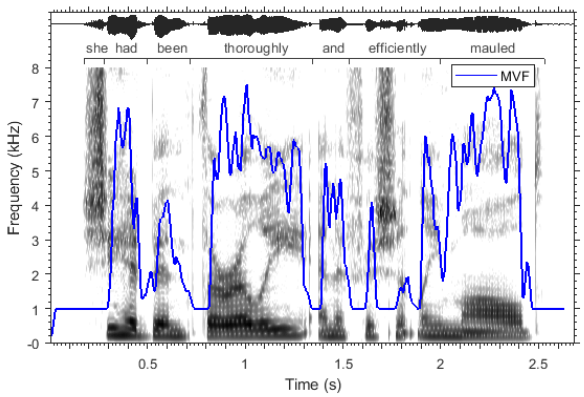

Figure 3: *Top: text and waveform, bottom: spectrogram and MVF of the synthesized speech from a female speaker.*

Table 1: *Parameters type of applied vocoders.*

Vocoder	Parameter per frame
Continuous	F0: 1 + MVF: 1 + MGC: 24
WORLD	F0: 1 + Band aperiodicity: 5 + MGC: 60
STRAIGHT	F0: 1 + Aperiodicity: 1024 + Spectrum: 1024