INTERSPEECH 2019: Show & Tell Contribution
September 15–19, 2019, Graz, Austria

Web-Based Speech Synthesis Editor

Martin Grůber[1], *Jakub Vít*[2], *Jindřich Matoušek*[1,2]

[1]New Technologies for the Information Society, University of West Bohemia, Czech Republic
[2]Dept. of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Czech Republic
gruber@ntis.zcu.cz, jvit@kky.zcu.cz, jmatouse@kky.zcu.cz

Abstract

This paper presents a web-based GUI frontend for a backend TTS system, including an editor of the synthesized speech. The tool allows synthesizing speech from general texts using all available synthesis methods with both modifications within the speech synthesis process and subsequent modifications of the synthesized speech targeting for instance speech prolongation, shortening, pitch or volume increasing or decreasing, etc.
Index Terms: speech synthesis, speech modification, speech editor

1. Introduction

Speech synthesis has become very popular and is used in many practical applications, including e.g. announcements in shops or public transport stations, conversion of legal documents to a spoken form, or teaching foreign languages in e-learning software. For all these mentioned applications, there is usually no need to have the synthesized speech available on-line and immediately as the phrases to be spoken are known in advance and are mostly static (not changing much in time). Instead, the requirement is that the synthesized speech is of a very high quality, without any glitches and distortions, with natural prosody, suitable for the environment it is going to be used/played in, and possibly fitting pre-defined time slots. Thus, the spoken (synthetic) phrases are usually prepared in advance, tested and evaluated, and tuned so that they sound as good as possible for that specific purpose. This kind of speech synthesis is usually not intended for ordinary end users but rather for companies or institutions providing services in the aforementioned areas in cases when a human speaker is not available for recording the requested phrases beforehand, or if it would be very expensive to hire one for this task. In addition, using a TTS system is more flexible as e.g. new phrases can be easily added or modified without a need to hire the same human speaker again or without a risk to mix different voices within a single service.

To make the tuning of the synthesized speech easier and more user-friendly, a tool containing a speech synthesizer and an editor of the synthesized speech was developed. It allows synthesizing speech from general texts in various languages and voices using different synthesis methods and mainly influencing the process of speech synthesis (depending on the method) in a way that the resulting synthetic speech is subjectively of the highest possible quality without audible artefacts fitting the target environment.

For example, when using the unit selection method for speech synthesis, glitches in prosody may occur in the synthetic speech. This tool allows resynthesizing such a sentence using a different sequence of units – non-optimal in terms of an objective criterion of the unit selection mechanism but optimal in terms of a subjective evaluation. Or, a synthesized announcement is supposed to fit in a defined time slot. This tool allows speeding up the speech or just some parts of it (either by influencing the speech synthesis process or by post-processing the resulting synthetic speech) so that its length is as requested.

The tool is divided in two parts: the synthesizer, which is used for speech synthesis, is described in Section 2, and the editor, which is used to influence the synthesis process and to do other requested modifications of the resulting speech signal, is described in Section 3.

2. Synthesizer

The synthesizer part of the tool is a GUI frontend for our TTS system ARTIC [1]. The synthesizer allows synthesizing speech from general texts using all available speech synthesis methods and voices. Currently, our TTS system provides two speech synthesis methods to be used by this tool: unit selection (using large speech corpora) and statistical parametric synthesis (further referred to as SPS; using neural networks) [2].

Unit selection method is a mature well-known technique used for TTS. Very briefly, it concatenates speech units (small segments of speech) from a large speech corpus (of a single speaker) in a way that the produced sequence of units meets some pre-defined criteria evaluating both the appropriateness of units in particular contexts and the smoothness in concatenation points of the units' speech signal.

SPS has become more popular in recent years due to the growing interest in using deep neural networks (DNNs) for this purpose. In this tool, two types of SPS are available. Both are using DNNs for generating speech parameters but different vocoders are used to generate speech samples. First one is using the WORLD vocoder [3] producing speech of a bit lower quality (still sufficient for some purposes) but fast enough whereas the second one using WaveRNN [4] is generating speech of a very high quality but it is much slower.

For both synthesis methods, various voices for various languages (Czech, Slovak, English, Russian, Armenian) are available in our system.

3. Editor

The editor allows modifications of the speech synthesized by the synthesizer. Depending on the synthesis method, various modifications are available. A screenshot of the editor is depicted in Figure 1. At the top of the page, there is the sentence requested to be synthesized. Beneath that, a phonetic transcription in a phonetic alphabet is displayed. The largest part of the screenshot is occupied by the waveform with the speech signal. At the bottom of the page, the F0 contour of the synthesized utterance is shown (if appropriate for that particular synthesis method).

In addition, at the very top of the page there is a menu allowing users to do some operations (described further) with the synthesized utterance. Users can also select one unit (phone,

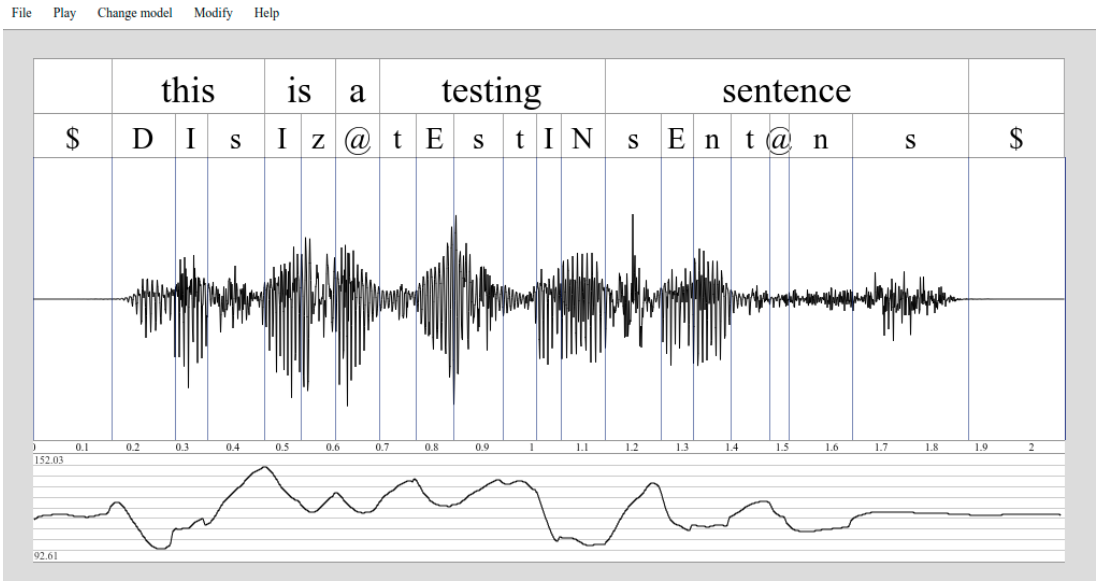

Figure 1: *Screenshot of the editor GUI.*

diphone, etc., depending on the synthesis method), more consecutive units, or a whole word and apply the operations only to the selection.

For the unit selection method, there is the *Change units* menu item. The user can e.g. re-synthesize the sentence (or just the selection) using different units, i.e. the best sequence of units (in terms of unit selection criteria) will not be used but another sequence (the next best one) will be composed instead. This might be useful if the original synthesized utterance contains some kind of distortions or artefacts. This way the user might select units within the non-naturally sounding part of the utterance and get another variant, not optimal in terms of unit selection criteria, but possibly avoiding the artefacts in the original variant and thus optimal in terms of a subjective evaluation.

Next, the user can change the criteria that drive the unit selection mechanism. The user might want to increase/decrease the speed/pitch of the synthesized utterance or the selected part of it. By changing these criteria, different units meeting these new requirements better might be selected if available in the speech corpus. If no other appropriate units can be found, no change is done and the user is informed about it. There are no signal modifications applied in this phase.

For the SPS method, the user can change the speed and pitch using the *Change model* menu item. These operations modify the underlying model characteristics that is further used for generating the speech samples. Thus, the user might directly affect the way the sentence is synthesized. By default, the speech is generated using the WORLD vocoder as it is fast. If the user is satisfied with the model settings, the utterance can be re-synthesized using the WaveRNN vocoder to achieve higher quality. This re-synthesis takes more time to finish as the WaveRNN vocoder works slower.

In addition, for both synthesis methods, the user might further change the resulting utterance characteristic by modifying the speed, pitch and volume using a post-processing tool (the *Modify* menu item). However, this operation modifies directly the speech signal and thus might deteriorate the quality of the synthesized speech.

4. Conclusion

This paper presents an application allowing "interactive speech synthesis", i.e. synthesizing texts using various speech synthesis methods and then interactively modifying both the speech synthesis process itself and also the resulting synthetic speech via a post-processing tool. The main purpose of this tool is to allow fine-tuning of the result of the speech synthesis process and thus improving the quality of the synthetic speech if the synthesized utterances can be prepared in advance and on-line synthesis is not required.

5. Acknowledgements

This research was supported by the Technology Agency of the Czech Republic, project No. TH02010307, by the Ministry of Education, Youth and Sports of the Czech Republic, project No. LO1506 and by the grant of the University of West Bohemia, project No. SGS-2019-027.

6. References

[1] D. Tihelka, Z. Hanzlíček, M. Jůzová, J. Vít, J. Matoušek, and M. Grůber, "Current state of text-to-speech system ARTIC: A decade of research on the field of speech technologies," in *Text, Speech, and Dialogue*, Cham: Springer International Publishing, 2018, pp. 369–378.

[2] J. Vít, Z. Hanzlíček, and J. Matoušek, "On the analysis of training data for wavenet-based speech synthesis," in *ICASSP*, Calgary, Canada, 2018, pp. 5684–5688.

[3] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[4] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *CoRR*, vol. abs/1802.08435, 2018. arXiv: 1802.08435.