Human Language Technologies – The Baltic Perspective
K. Muischnek and K. Müürisep (Eds.)

doi:10.3233/978-1-61499-912-6-26

Towards a Modern Text-to-Speech System for Latvian

Roberts DARĢIS[1] and Ilze AUZIŅA
Institute of Mathematics and Computer Science, University of Latvia

Abstract. Thanks to the advancements in neural networks there have been many new breakthroughs in human-like speech synthesis over the last couple of years. For Latvian, there have been no new publications about speech synthesis since 2010. The paper describes efforts to apply recent advancements in neural speech synthesis to Latvian using open-source tools.

Keywords. speech synthesis, text-to-speech system, neural network models

1. Introduction

Speech synthesis as a topic in natural language processing has regained its popularity after DeepMind [1], Baidu [2] and Google [3] published their latest results in human-like speech synthesis. Although the code itself was not made public, many have tried to reproduce their result with various success.

For Latvian, there are two text-to-speech systems published, both using the concatenative speech synthesis method [4][5]. Since then methods for more natural sounding voice synthesis are available. This paper describes a recent attempt in low-cost yet qualitative text-to-speech synthesis for Latvian. The aim of experiments conducted in this article is to find out how human-like voice can be achieved using open-source text-to-speech development systems.

2. Specifics of Latvian

The Latvian language is spoken by 2 million people (including non-native speakers) and it is the only official language in Latvia and one of the working languages of European Union.

Although Latvian is a language with a relatively simple relationship between orthography and phonology it has several specific properties that affect speech synthesis:

- Vowel quantity is phonemic and plays an important role in the language (pile ‘drop’, pīle ‘duck’).
- Syllable rhythm and syllable intonation (pitch inflections on syllables) determines the long syllable pronunciation.

[1] Corresponding author; Artificial Intelligence Laboratory, Institute of Mathematics and Computer Science, University of Latvia, Raina blvd. 29, Riga, LV-1459, Latvia; E-mail: roberts.dargis@lumii.lv.

- Some phonological processes affect the pronunciation of words.

These properties have to be taken into account in text normalization, grapheme-to-phoneme modeling and prosody generation.

Although text normalization has a huge impact on the audio understandability and the overall experience, this is not done in the context of this article, because the aim of experiments conducted in this article was to find out how human-like voice can be achieved using open-source text-to-speech building systems.

An improved version of the previously developed rules-based system [5] was used for grapheme-to-phoneme modeling.

3. Corpus for speech synthesis

As shown in Wavenet [1], Deep voice [2] and Tacotron [3] text-to-speech systems, 30 hours of training data are enough to create a human-like state-of-the-art artificial voice.

For Latvian, there is no publicly available corpus for text-to-speech synthesis. Latvian Speech Recognition Corpus [6] is the closest available corpus for speech synthesis. Unfortunately, there are multiple obstacles for using this corpus to train text-to-speech system. First, no speaker contains more than 1 hour of recordings in the corpus. Second, only 38 out of 100 hours contain prepared speech, other 62 hours are spontaneous speech that in most cases are not viable for the training of a text-to-speech system. A part of the corpus contains background noises as well.

Recording a new corpus requires a professional orator and transcribing an existing audio is time-consuming so alternative, less human-resource demanding options were chosen. Two different corpora were created for text-to-speech experiments: corpus based on audiobooks and corpus based on news.

The corpus based on audiobooks was created from fairy tales, using automatic speech recognition for alignment. Almost 4 hours of aligned text and audio was obtained to be used as a training corpus.

The corpus based on the news was created in completely unsupervised manner, using automatic speech recognition and speaker diarization, so the corpus might contain segments from different speakers and the text might not be completely accurate. Almost 30 hours of training data were obtained from this approach.

4. Text-to-speech system

Experiments with two different text-to-speech systems were conducted:

- Merlin toolkit [2] for building Deep Neural Network models for statistical parametric speech synthesis [8].
- Deepvoice3_pythorch toolkit [3] - a community-based implementation of methodology published by other authors [2][9] for training convolutional sequence-to-sequence model with attention for text-to-speech synthesis.

Training of the text-to-speech system on Merlin toolkit on single GPU took about 6 hours for audiobooks corpus and about 24 hours for the news corpus and the synthesized voices are understandable.

[2] Merlin Github repository: https://github.com/CSTR-Edinburgh/Merlin

[3] Deepvoice3_pythorch Github repository: https://github.com/r9y9/deepvoice3_pytorch

The experiments with Deepvoice3_pythorch toolkit was inconclusive. The training of English text-to-speech system based on The LJ Speech Dataset[4] worked fine. The training of Latvian text-to-speech system based on audiobook corpus failed due to undetermined technical errors. There were no technical difficulties in the training of Latvian text-to-speech system based on news corpus, but after 50k iterations the generated audio was not understandable, although for the English text-to-speech system the audio was understandable after 20k iterations. As mentioned in Deep Voice 3 paper [2], the system they trained converged after 500k iterations, so there is still possibility that voice will become understandable and even maybe human-like. Further experiments are required to find out the causation of the problems and to determine if this implementation is applicable for development of Latvian human-like artificial voice.

In conclusion, two voices were trained from audiobooks corpus and news corpus, using Merlin toolkit for building Deep Neural Network models for statistical parametric speech synthesis that will be used in the further evaluation.

5. Evaluation

To compare the two newly created systems and the concatenative speech synthesis method developed in 2010 [5] three samples were generated, where each sample should give some advantage to one of the systems:

- A few sentences from the weather forecast, because the concatenative system was developed from weather forecasts.
- A few sentences from a fairytale, because one of the new systems was trained on fairytale audiobooks.
- A few sentences from the news, because one of the new systems was trained on news.

All of the 9 samples are available online[5] (three different sentences from three different system) and was evaluated by 10 different people independently, giving each sample an opinion score in scale for 1 to 5 (from bad to excellent). Mean opinion score (MOS) were calculated and are shown in Table 1. As the evaluation shows, both of the newly trained systems are more preferred, compared to concatenative speech synthesis system developed previously. For comparison there are samples available online[6] in English from text-to-speech system trained on the same settings using CMU_ARCTIC database [10].

Comparing the newly developed systems, the system trained on the news got higher score mainly because of the difference in voice quality that could be explained by the huge difference in the size of the corpora (4 hours vs 30 hours). Conducting the post-evaluation interviews, reviewers admitted that although they preferred the system trained on the news, the biggest downside to that system compared to the audiobooks system was the lack of pause and the high tempo of pronunciation. The fact that the news corpus didn't have any punctuation marks and the segments were not divided into logical phrases might have played a big role in this. Experiments with a combination of both models might yield better results.

[4] The LJ Speech Dataset homepage: https://keithito.com/LJ-Speech-Dataset/

[5] The voice samples used in evaluation: http://runa.korpuss.lv/tts/

[6] Text-to-speech sampels in English: https://cstr-edinburgh.github.io/merlin/demo.html

Table 1. Mean opinion score (MOS) of the generated voice sample

	Sample 1	Sample 2	Sample 3	Sample 4
Concatenative	2.00	2.00	1.70	1.90
Audiobooks	3.20	2.80	1.90	2.63
News	3.80	4.00	3.60	3.80

6. Conclusion and further work

Multiple voices using neural network models have been created. Evaluation shows that people preferer newly created statistical parametric speech synthesis over previously developed concatenative speech synthesis, but there is still a long road ahead to achieve truly human-like speech synthesis for Latvian.

Further work includes experiments with different text-to-speech systems to find one that works the best and is most easily adaptable to achieve an even better result. Prosody generation will also be a task to consider when reasonable acoustic quality is achieved.

Text normalization is a complex task, especially in Latvian, because Latvian is a highly inflected language where endings of the replaced word must be matched with the context. Text normalization is not considered to be a priority until acoustically qualitative speech synthesis is obtained.

References

[1] Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. *Wavenet: A generative model for raw audio.* arXiv preprint arXiv:1609.03499 (2016).

[2] Arik, S. O., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., ... & Sengupta, S. *Deep voice: Real-time neural text-to-speech.* arXiv preprint arXiv:1702.07825 (2017).

[3] Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... & Le, Q. *Tacotron: Towards End-to-End Speech Synthesis.* arXiv preprint arXiv:1703.10135 (2017).

[4] Goba, K., & Vasiļjevs, A. (2007). *Development of text-to-speech system for latvian.*

[5] Pinnis, M., & Auziņa, I. (2010, August). *Latvian text-to-speech synthesizer*. In Proceedings of the 2010 conference on Human Language Technologies--The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT 2010 (pp. 69-72). IOS Press.

[6] Pinnis, M., Auzina, I., & Goba, K. (2014). *Designing the Latvian Speech Recognition Corpus.* In LREC (pp. 1547-1553).

[7] Ping, W., Peng, K., Gibiansky, A., Arik, S., Kannan, A., Narang, S., ... & Miller, J. (2018). *Deep voice 3: Scaling text-to-speech with convolutional sequence learning*. In Proc. 6th International Conference on Learning Representations.

[8] Wu, Z., Watts, O., & King, S. (2016). *Merlin: An open source neural network speech synthesis system.* Proc. SSW, Sunnyvale, USA.

[9] Tachibana, H., Uenoyama, K., & Aihara, S. (2017). *Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention*. arXiv preprint arXiv:1710.08969.

[10] Kominek, J., Black, A. W., & Ver, V. (2003). *CMU ARCTIC databases for speech synthesis*.