Instructions for ACL 2018 Proceedings

Dante Razo
Indiana University / Bloomington, IN
drazo@indiana.edu

Abstract

Chuvash is a minority language spoken by roughly one million people in European Russia (RBS, 2012). For this project, I trained an Ossian/Merlin speech synthesis model on Chuvash-language news clips. The former is a frontend for Merlin, which is a neural net based Speech Synthesis System.

This project aims to train neural net based speech synthesizers to produce intelligible Chuvash from text samples. The synthesizers used are Ossian, Mozilla TTS, and Mozilla LPCNet. At the minimum, the performance of the three will be compared. The Expected Product involves cleaning the Chuvash data to produce better results with one of the three synthesizers. If time permits, the best-performing synthesizer will be tweaked to produce even more accurate results. Data will come from reliable sources such as the Apertium project. The end goal is to train a neural network that mimics Chuvash speech to some degree, then improve on it as much as time allows.

1 Introduction

The Chuvash language is spoken by roughly one million people in European Russia (RBS, 2012). Despite the large number of speakers, it is considered a minority language. This project aims to train popular speech synthesizers to produce intelligible Chuvash from written samples.

Text-to-speech works by accepting text, conducting linguistic analysis on the input, then producing audio waveforms. Data preprocessing techniques include normalization and tokenization (Wikipedia, 2019a), though the latter is meant for text corpora used as input. Audio could be "preprocessed" by normalizing volume and editing silence out from samples.

The DNNs used in this project will be trained on audio samples mainly taken from Chuvash-language news programs. Due to the small number of samples ($\sim$ 546) for the task at hand, I'll likely need to do transfer learning to get the best results. This technique is explained in further detail in §2.3.

2 Speech Synthesis

Speech synthesis is the production of artificial human speech (Wikipedia, 2019b). Text-to-speech (TTS) is a subset of the field which focuses on taking text as input, and returning audio "speech" as output. There are multiple ways to do TTS, but this project will focus on Deep Neural Networks (DNNs). The best DNNs on the market sound like real humans, but they can still be distinguished by their staggered manner of speaking.

2.1 Deep Neural Net Systems

2.2 Data & Corpora

I used Francis Tyers' `Turkic_TTS` repository of Chuvash-language news clips to train my model.

3 Training

3.1 DNN Models

4 Results

4.1 Model Evaluation

License

This project is under the `GPL-3.0` license.

Acknowledgments

The acknowledgments should go immediately before the references. Do not number the acknowl-

edgments section (*i.e.*, use `\section*` instead of `\section`). Do not include this section when submitting your paper for review.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Russian Bureau of Statistics RBS. 2012. Population of the russian federation by languages (in russian).

Wikipedia. 2019a. Lexical analysis.

Wikipedia. 2019b. Speech synthesis.