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. The performance of Ossian is compared to other popular solutions such as Mozilla TTS, Mozilla LPCNet, Festival, and eSpeak. Further experiments include tweaking Ossian/Merlin to produce better results.

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.

3 Deep Neural Net Systems

DNN-based systems are among the best in speech synthesis, but is the difference that extreme? Let's take a look at the models used in this project.

3.1 Ossian & Merlin

3.1.1 Parameters

3.2 Mozilla TTS

3.3 Mozilla LPCNet

3.4 Festival

3.5 eSpeak

3.6 Data & Corpora

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

4 Training

4.1 Models

5 Results

5.1 Mozilla TTS

5.2 Mozilla LPCNet

5.3 Festival

5.4 eSpeak

5.5 Model Evaluation

5.5.1 Scoring System

6 Improving Models

6.1 Ossian & Merlin

License

Acknowledgments

I'd like to acknowledge Francis M. Tyers' work on the `Turkic_TTS`

References

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

Wikipedia. 2019a. Lexical analysis.

Wikipedia. 2019b. Speech synthesis.