Deep Neural Net Speech Synthesis for the Chuvash Language

Dante Razo
Indiana University, Bloomington, IN
Department of Linguistics
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

Despite the relatively large number of speakers, Chuvash is still considered a minority language. This project aimed 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 systems used in this project will be trained on audio samples mainly taken from Chuvash-language news programs. Due to the small number of samples (~ 546) for the task at hand, transfer needed should be used to get the best results. Unfortunately, given the time restraint, I was unable to implement transfer learning with Ossian or the other systems.

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 Neural Networks, with one case of formant synthesis. The best neural nets on the market sound like real humans, but they can still be distinguished by their staggered manner of speaking.

3 Text-to-Speech Systems

Neural net based systems are among the best in speech synthesis, but is the difference noticeable to the layman? Let's take a look at an overview of the models used in this project:

System	Basis	Score
Ossian/Merlin	DNN	*TBD*
Mozilla LPCNet	RNN	*TBD*
Festival	*TBD*	*TBD*
eSpeakNG	Formants	*TBD*

Table 1: Speech synthesis systems.

eSpeakNG is the black sheep of the bunch; it uses *formant synthesis* to create language models.

3.1 Ossian & Merlin

Ossian is a front-end for speech synthesis development by the Centre for Speech Technology Research (CSTR) at The University of Edinburgh. It's meant to interface with the Merlin library for neural-net based speech synthesis developed by the same team.

3.1.1 Parameters

3.2 Mozilla LPCNet

3.3 Festival

Modular, TTS

3.4 eSpeakNG

Formant synthesis

3.5 Data & Corpora

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

1. Turkic_TTS by Francis M. Tyers (ftyers)
2. apertium-chv (GPL-3.0) by Apertium

3.5.1 Repositories

1. Mozilla TTS (MPL-2.0) by Mozilla
2. Ossian (Apache-2.0) by CSTR-Edinburgh
3. Mozilla LPCNet (BSD-3-Clause) by Mozilla

4 Training

4.1 Models

5 Results

5.1 Ossian & Merlin

5.2 Mozilla LPCNet

5.3 Festival

5.4 eSpeakNG

6 Model Evaluation

6.1 Scoring System

7 Improving Models

7.1 Ossian & Merlin

License

Acknowledgments

I'd like to acknowledge Francis M. Tyers' work on the `Turkic_TTS` corpus. Compiling the data was easy using the provided instructions. I'd like the thank Professor Tyers again for his website, reading, and tutorial recommendations. Thanks to the original authors of the ACL 2018 format as well for this template. Finally, thanks to Josh Meyer for his extremely helpful **Ossian** and **eSpeakNG** tutorials.

References

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

Wikipedia. 2019a. Lexical analysis.

Wikipedia. 2019b. Speech synthesis.