Developing a Kyrgyz Speech Synthesizer

A Demonstration of Ossian, Merlin, and Kaldi toolkits

Joshua Meyer

jrmeyer.github.io

@joshmeyerphd

October 18, 2017

The Problem

So you want to make a synthesizer?

Without Data	
Approach:	rule-based
Required:	linguist
Pros:	no data, can speed-up
Cons:	robotic sounding
Example:	eSpeak NG

With Data	
Approach:	statistical
Required:	speech corpus
Pros:	human-like speech
Cons:	data == $$$
Example:	Merlin

Overview

Development Pipeline

1. find audiobook (Human)
2. split audiobook on silence (Human)
3. hand-align sample of audiobook (Human)
4. train speech recognizer on sample of audiobook (Kaldi)
5. use new recognizer to generate transcripts for more audiobook (Kaldi)
6. use text from audiobook to train TTS frontend (Ossian)
7. train acoustic and duration model for DNN TTS (Merlin)
8. synthesize new speech (Merlin / Ossian)

DATA

DATA

Audiobooks

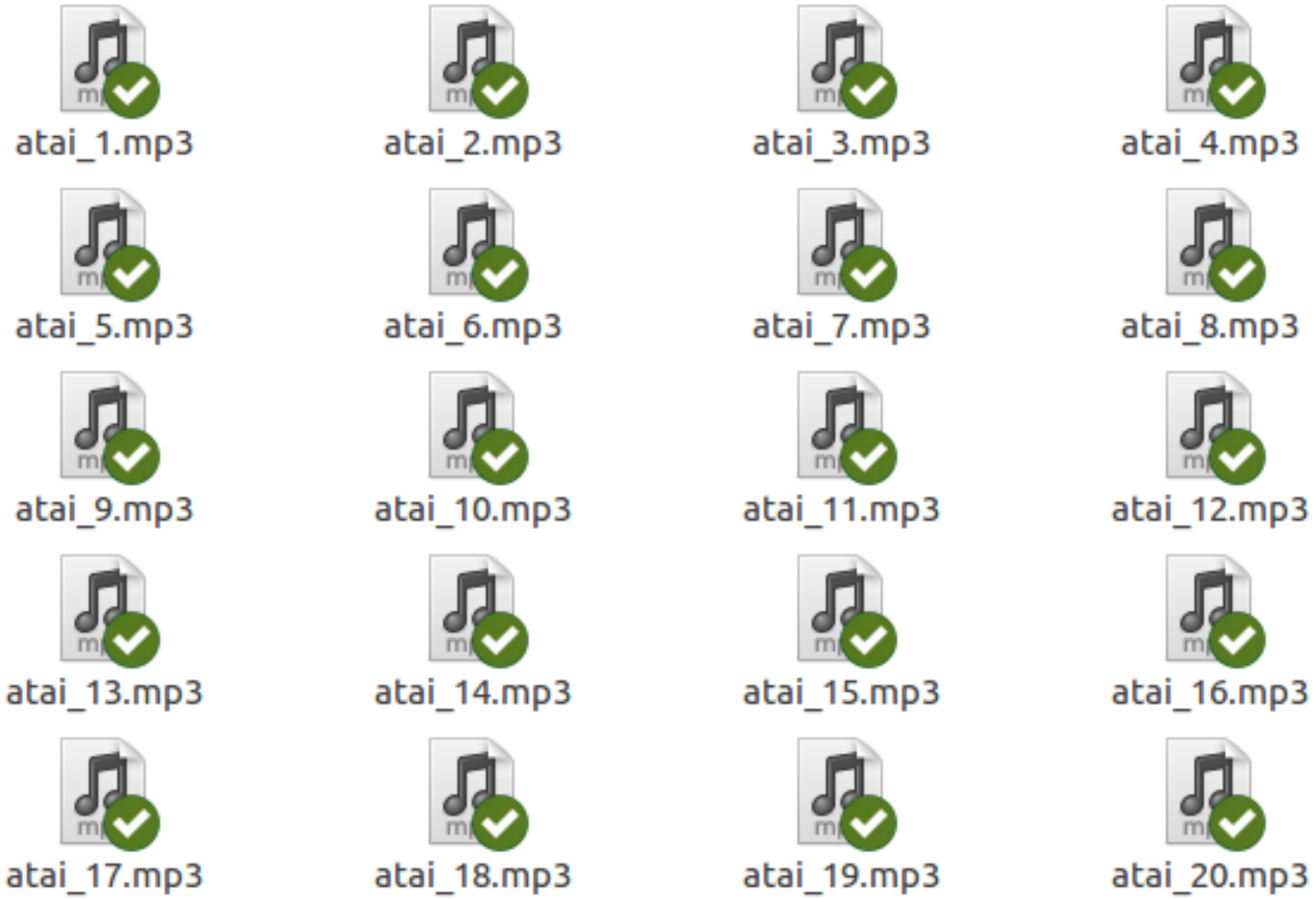
More data == better.

Better data == better.

Kyrgyz books == `bizdin.kg`

Audiobooks

Original Audio



Original Text

К.Каимов.

АТАЙ

Зарыктырган үмүт

Мобу, жүк тактайдай тептегиз болуп, кырка тарткан жепирекей дөңсөөнүн алдындагы алакандай жайык Көк-Кашат деп аталат. Мунун дарегин балким ушул айылдан башка жактагылар билишпес. Мындай көз жаздымында калып, учуру келгенде гана эске алынчу жайлар дүйнөдө мол го. Ага бет маңдайлаш күңгөй тарапта, атактуу Кең-Колдун оозундагы дарбазадай кызыл таш, айкөл Манастын күмбөзү тарыхтын күбөсүнө окшоп, бул кай жер экенин өзү эле айкындайт.

Өрөөндүн ортосунда күкүктөп агып жаткан чоң сууну бойлой өскөн чытырман токой да арстандын жалына окшоп дүпүйүп, баатырдын жеринин элесин көз алдыга тартат. Күңгөй-тескей жактары ажыдаардын азуусундай арсак тоолор менен курчалган кең мейкин чыгыштан батышты көздөй тасырайып созулуп жатат.

Split & Align

Split audio on silence: `sox` or `Audacity`

Split text on utterance-like punctuation: python3

Align text to audio: `ears, eyes, hands`

Split & Align

Train Speech Recognizer

Language model == audiobook text.

Overfit to speaker == good.

Train w/ `Kaldi`.

Train Recog

Decode Speech

Split on silence **exactly** as before.

Decode w/ `Kaldi`.

Now you have **more training data** for TTS.

Decode Spee

TTS

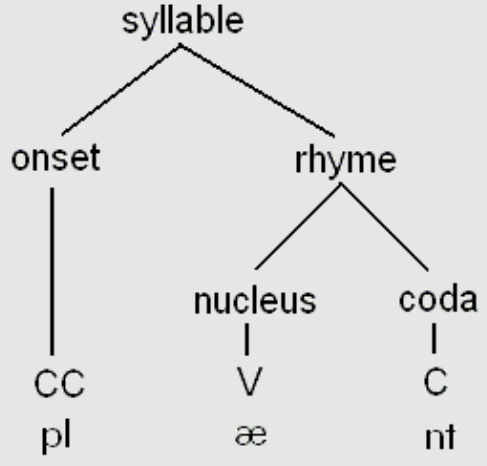
TTS

Train Ossian Frontend

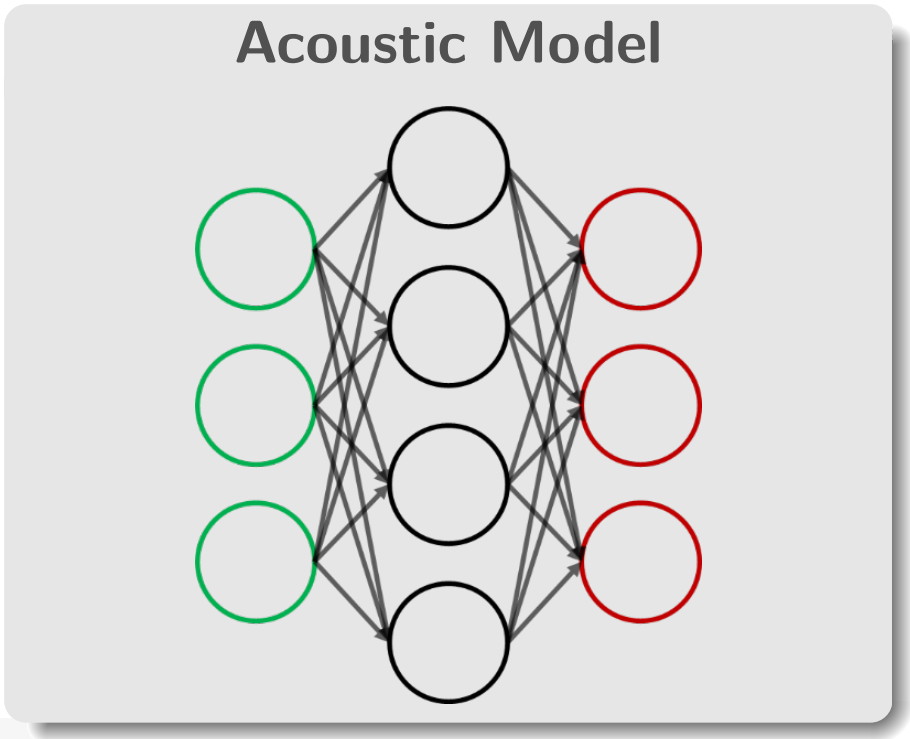
Naive Approach

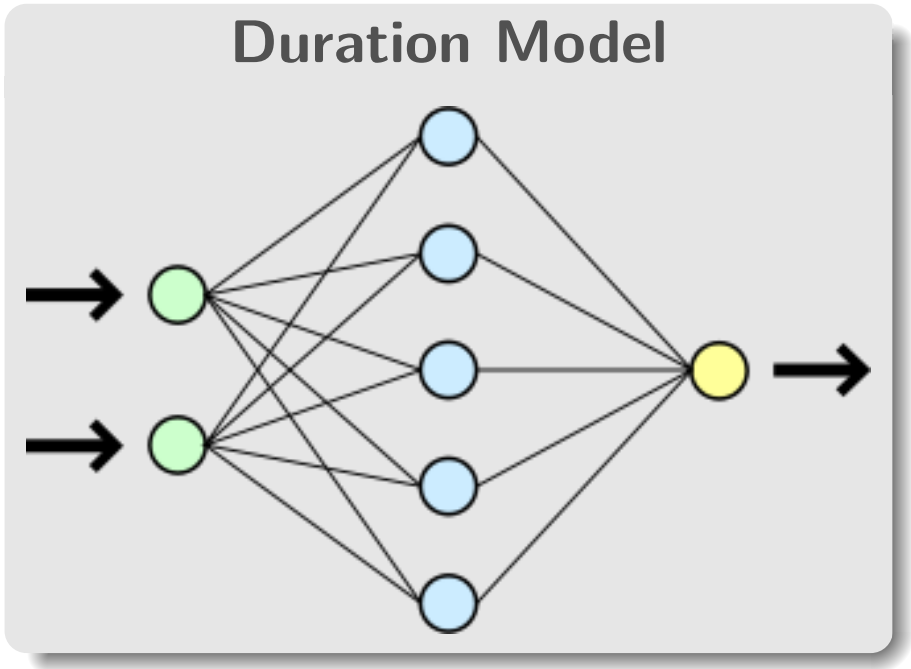
`tokenize` on whitespace
`word2vec` (instead of `POS`)
split into `characters`

Linguistic Approach



Train Merlin Models





Synthesize New Speech

trained model + new text == new speech

simple as that

Synthesize

Thanks!

Чоң рахмат!

Рахмет!

Күп рәхмәт!

Thank you!

Acknowledgements

I would like to thank the researchers at CSTR for welcoming me to Edinburgh and helping me understand their fabulous toolkits; in particular Oliver Watts, Simon King, and Srikanth Ronanki.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. (DGE-1746060). Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation.

Thanks!