Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5B
D-70569 Stuttgart

Master thesis

Phonetic Representations of Speech for Human Pronunciation Feedback and Automatic Accent Transfer

Isaac Riley

Studiengang:	M.Sc. Computational Linguistics

Prüfer*innen:	Prof. Dr. Wolfgang Wokurek
	Prof. Dr. Antje Schweitzer
Betreuer:	Prof. Dr. Wolfgang Wokurek

Beginn der Arbeit:	01.04.2021
Ende der Arbeit:	01.10.2021

Erklärung (Statement of Authorship)

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst habe und dabei keine andere als die angegebene Literatur verwendet habe. Alle Zitate und sinngemäßen Entlehnungen sind als solche unter genauer Angabe der Quelle gekennzeichnet. Die eingereichte Arbeit ist weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens gewesen. Sie ist weder vollständig noch in Teilen bereits veröffentlicht. Die beigefügte elektronische Version stimmt mit dem Druckexemplar überein.[1]

(Isaac Riley)

[1] Non-binding translation for convenience: This thesis is the result of my own independent work, and any material from work of others which is used either verbatim or indirectly in the text is credited to the author including details about the exact source in the text. This work has not been part of any other previous examination, neither completely nor in parts. It has neither completely nor partially been published before. The submitted electronic version is identical to this print version.

Acknowledgments

XXX

Inhaltsverzeichnis

1 Introduction

In linguistics, the notion of accent refers to a pattern of pronunciation that does not change the semantic content of what is uttered, but which may carry pragmatic meaning and convey demographic information about the speaker. While some phonetic variation may be random or occur at the individual level (idiolect), accent typically varies systematically according to native language and dialect, which in turn vary geographically and demographically.

Accent is not usually an impediment to human understanding of speech; this is especially true among native speakers. In automatic speech recognition, nonstandard accent typically has a much stronger negative effect on recognition accuracy. For this reason, accent is an active subject of research in automatic speech processing.

Related to accent transfer is the task of voice conversion. The goal of voice conversion is to generate speech containing identical linguistic information as an input sample, modifying only the timbre to match a target speaker.

2 Background

2.1 Automatic Speech Recognition

Automatic speech recognition

2.2 Text-to-Speech Synthesis

2.3 Speech Signal Processing

2.3.1 Short-Time Fourier Transform

2.3.2 Mel Spectrum and Coefficients

2.3.3

2.3.4

2.4 Neural Networks

2.4.1 Feedforward Neural Networks

A feedforward neural network

2.4.2 Recurrent Neural Networks

Feedforward neural networks are effective in processing vector inputs. If the input data

One powerful and popular architecture for recurrent neural networks is the long short-term memory (LSTM) network.

2.4.3 Convolutional Neural Networks

When input data have a spatial component, it is beneficial to use a network architecture that

2.4.4 Generative Adversarial Networks

Operating within the parradigm of supervised learning, neural networks can be used for classification by learning a posterior distribution over labels, given numerical features representing features of an input sample. Given sufficient exposure to labeled training data, an appropriately designed model can update its parameters so as to assign an increasingly larger probability to the correct label.

Another branch of machine learning involves generative learning. Given a training dataset, a generative model learns to output samples resembling those from the training dataset.

In “vanilla” generative adversarial networks (GANs), the are two networks whose training takes the form of a minimax game, in which each model seeks to maximize the loss of the other. The first network is the generator, which is given random noise as input and tasked with outputting samples of the same type as the target data. The second model is the discriminator, a classification network tasked with classifying generated samples as real or generated. Throughout the (ideal) adversarial training process, the generator

3 Related Work

3.1 Models Used

3.1.1 Tacotron 2

Tacotron 2 is a deep neural sequence-to-sequence model that generates mel spectrograms from text. I can be combined with a deep neural vocoder, it forms an end-to-end text-to-speech system.

3.1.2 WaveGlow

WaveGlow is a deep neural flow-based vocoder that generates waveform audio from mel spectrograms.

3.1.3 Allosaurus

3.1.4 CycleGAN

In ‘vanilla’ GANs, the input is typically random noise, which is mapped to some point in the target distribution. CycleGAN

3.1.5 SeqGAN

4 Resources

5 Methods

5.1 Phonetic Representation-Based Speech Synthesis and Accent Transfer

5.1.1 Speech Synthesis from Phonetic Representations

Text here.

5.1.2 Accent Transfer in Phonetic Representation Space

Text here.

5.2 Phonetic Transcription-Based Speech Synthesis and Accent Transfer

5.2.1 Transcription

5.2.2 Rule-Based Segmental Accent Transfer

Because accents tend to be characterized by a number of salient features differing from the 'standard' dialect or simply from other dialects, it is possible to formulate mapping rules that define which phonetic changes must be made to a source accent to make it sound more like a target accent. For example, a speaker of Received Pronunciation English wishing to imitate a speaker of American English will typically pronounced syllable-final "*r*" as [ɹ] and will voice intervocalic occurrences of "*t*". An American English speaker wishing to imitate RP English might apply the inverse of these rules.

The following section seeks to investigate and formalize approaches to learning phonetic transformation rules from non-parallel data.

5.2.3 GAN-Based Segmental Accent Transfer

6 Results

7 Conclusions

Literatur