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

Master thesis

Representational and Segmental Approaches to Accent Conversion

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:	06.04.2021
Ende der Arbeit:	06.10.2021

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

I wish to express my thanks to Dr. Wokurek for his mentoring and encouragement. He was incredibly generous with his time, and his advice and willingness to listen to my ideas helped me to keep a level head throughout what was at times a difficult research process. I learned from him more than he may suspect, and I will carry his insights and advice with me long after I have submitted my thesis.

I am eternally indebted to my wife, Sandra, for her assistance and encouragement. In so many ways, she made this possible. To her I owe not only my thesis, but also my sanity.

Contents

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 occasionally an impediment to human understanding of speech; however, among native and advanced speakers, accent is often no impediment to communication. In automatic speech recognition, accent accents underrepresented in training datasets typically have 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. In some respects, voice conversion is a simpler task; namely, there is no temporal realignment and phonetic information remains the same. The absence of temporal realignment means that voice conversion models have an equal number of frames in input and output. This makes possible frame-to-frame mappings, also surrounding context must still be taken into account in order to achieve good results.

This works explores approaches to accent conversion with and without voice conversion. These approaches may be referred respectively as intra-speaker accent conversion and cross-speaker accent conversion. In intra-speaker accent conversion (henceforth simply *accent conversion*), an utterance from the source speaker is modified to match the accent of the target speaker, without changing voice quality or non-accent-carrying linguistic content. By contrast, in cross-speaker accent conversion, accent conversion is combined with voice conversion. Henceforth, this will simply be referred to as speaker conversion.

This work investigates two distinct but related approaches to speech conversion,

both accent conversion and speaker conversion. The first is a continuous approach in which speech is mapped to and synthesized from a continuous phonetic representation. The second is a discrete approach, in which speech is transcribed to and synthesized from the International Phonetic Alphabet (IPA).

Clearly, in the simplest case, speaker conversion may involve the combination of ASR and TTS, where the ASR model recognizes the speech of the source speaker and the TTS model synthesizes the speech of the target speaker from that same text. This work takes an alternative approach, in which speech is transcribed in the IPA, modified by an intermediate model to match the phonetic patterns of the target speaker, and a TTS model trained to synthesize speech from IPA is used to synthesize the desired speech.

Summary and Comparison of Speech Conversion Tasks

Speech Aspect	Accent Conversion	Voice Conversion	Speaker Conversion
content	source	source	source
accent	**target**	source	**target**
vocal quality	source	**target**	**target**

2 Related Work

2.1 Prior Work on Accent Transfer

Much of the work in accent conversion has been focused on foreign accent conversion, in which a non-native accent is modified to more closely resemble a native accent. This has obvious practical applications, but is also important in automatic speech recognition, since the best-performing models are typically trained on high-resource accents.

All approaches can be divided into parallel and non-parallel approaches, as in the related setting of voice conversion. In parallel accent conversion, data is available which contain identical utterances from source and target speakers, which in

non-parallel accent, the content of the utterance need not coincide. There are many advantages inherent in parallel approaches, but as parallel data is difficult to collect, it is often less practical. Non-parallel data is abundant, but presents another set of challenges, namely the inability to use supervised learning approaches. In other words, for a given source utterance, there is no correspondending 'label' indicating how that same utterance should sound following conversion. Thus, non-parallel approaches are forced to rely on distribution-matching techniques.

One simple approach is to combine the source and target utterance in some way. While somewhat crude in some ways, and arguably not 'pure' accent conversion as defined above, it nevertheless succeeds in reducing perceived foreign accent. However, the output contains characteristics of both the source and target utterance (Aryal et al., 2013).

A non-parallel approach is to map frames in the respective utterances into a common phonetic space using an aoustic model and then identify the frames of the target speaker that are phonetically the most similar to the reference source frame, using some similarity metric such as Kullback-Leibler divergence. This is an elegant and efficient approach; however, it suffers from a number of drawbacks. It is unable to handle differences in length between source and desired target, e.g. when a segment or sequence of segments typically occupies more frames in one speaker than another. It also constrains the target space to phones that are in fact realized by the target speaker, which may result in partial accent conversion. Despite these limitations, frame pairing has achieved some success and is a good baseline approach (Zhao and Gutierrez-Osuna, 2019; Zhao et al., 2018).

The third of the popular parallel approaches involves measurements from an electromagnetic articulograph. These are mapped to articulatory features, typically including MFCCs, using a GMM. The acoustic features are then normalized and provided as input to a appropriate vocoder such as STRAIGHT. (Aryal and Gutierrez-Osuna, 2014). However, articulatory approaches are generally found to be inferior to acoustic-based approaches, which is due in part to the fact that acoustic methods provide a more fine-grained representation with which to work. (Aryal and Gutierrez-Osuna, 2016).

Another class of accent conversion methods involves those in which speech synthesis plays the primary role, typically by representing speech in a shared intermediate representation and using a synthesizer trained on speech data of the target speaker or accent. One of these, Zhao et al. (2019), will be discussed below in greater detail.

Another synthesis-based accent conversion model is Parrotron, which makes heavy use of both ASR and TTS models. Designed to make human communication as well as the use of ASR systems easier for individuals with speech impairments or strong accents, Parrotron uses transfer learning to adapt a pretrained , already robust ASR model to a particular speaker, learning the patterns of errors occurring in the individual's speech. This model is then able to transcribe the speech with reasonable accuracy, which is then synthesized using a pretrained speech synthesis model. While this does not preserve voice quality, it does solve the problem of low intelligibility Biadsy et al. (2019).

Voice conversion is task that is in some ways closely related to accent conversion. In a sense, it may be thought of as a reverse, in that accent conversion generates an utterance with the same voice as and a different accent than the source utterance. In contrast to this, voice conversion generates an utterance with the same accent and different voice. Voice conversion is traditionally nearly always performed on mel spectrograms, although some progress has been made on waveform-to-waveform conversion.

2.2 Relevant Models

2.2.1 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.

2.2.2 WaveGlow

WaveGlow is a deep neural vocoder that generates waveform audio from mel spectrograms (Prenger et al., 2019). It has a flow-based architecture that learns an invertible mapping, converting a noise vector and additional conditioning input into nearly human-sounding waveform audio. Its audio generation is accomplished by sampling from a distribution, which is transformed as it passes through the model and is conditioned on mel-spectrograms representing the audio that is to be output in waveform.

It is based on the Glow architecture for image generation (Kingma and Dhariwal, 2018) and on WaveNet, the groundbreaking autoregressive waveform generation model (van den Oord et al., 2016). Its primary advantage over WaveNet is faster inference speed; because it is non-autoregressive, it can generate audio much faster than WaveNet, and 25 times faster than real time on a V100 GPU [XX].

2.2.3 CycleGAN

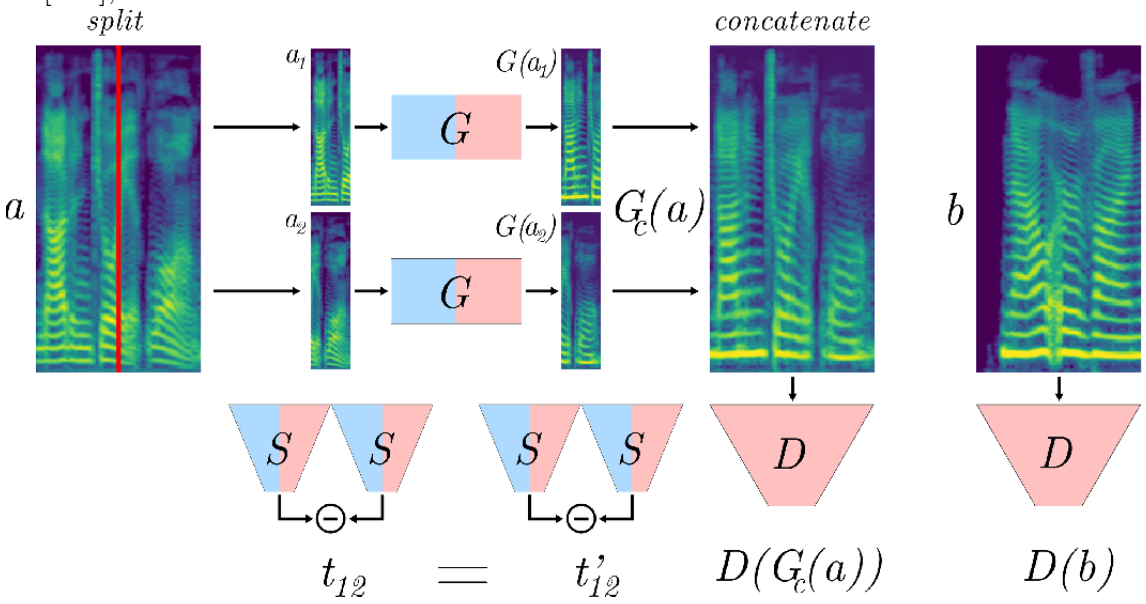
As described above, in 'vanilla' Generative Adversarial Networks, the input is typically random noise, which is mapped to some point in the target distribution. Moreover, because the generator is rewarded simply for generating some output appearing to have been drawn from the target distribution, there is no direct control over which point from the target distribution is generated. However, there are many applications in which one might wish to exert more control over the output by conditioning on some desired criteria. One common case involves the task of image translation, in which certain aspects of an image are altered while others remain intact. A simple example of image translation is a model that transforms horses into zebras. In this case, a simple GAN may take a horse image as input and output a very dissimilar zebra image as output; this is because there is no constraint to enforce content preservation. CycleGAN (Kaneko and Kameoka, 2017) achieves this by training an inverse generator to map the transformed image back to the original. This cyclical consistency constraint requires the forward generator to preserve semantic content from the original image so that it the original can be reconstructed

from the transformed image.

2.2.4 MelGAN-VC

Most work on GANs focuses on data with a fixed output shape, where input size and output shape are typically equal, such as in the case of image generation. When an input is supplied, it is for the purpose of conditioning the output and the input typically has the same shape as the output.

In settings where the data has an intrinsically sequential nature and the input size varies, it is not possible to require a fixed and shape. One such setting is voice conversion, and one simple and effective solution has been proposed by Marco Pasini in [XX], illustrated below:

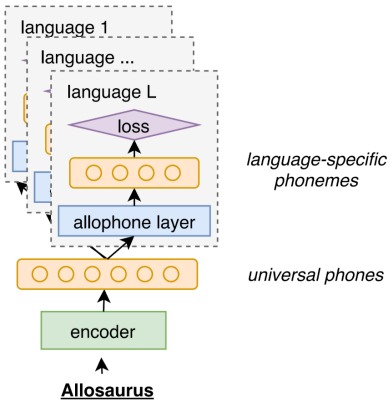


The key insight is that by taking a fixed length (T frames) of the input spectrogram, then splitting the input into two equal-sized blocks and passing each to the generator network. The converted spectrograms, i.e. the two generator outputs are then recombined and passed to the discriminator. Due to the adversarial loss training objective, the generator must learn to convert each block in such a way that, when reassembled, it will be indistinguishable from a real sample from the target set. Because the target set does not contain spectrograms with discontinuities, such

a discontinuity at the splitting point would clearly identify a fake sample. Thus, the generator is incentivized to produce outputs which, when re-joined, do not contain such telltale discontinuities at the splitting points. Because this applies to any split point, it is possible to train a relatively small generator that can receive inputs of arbitrarily length one block at a time and generate the desired output.

2.2.5 Allosaurus

Allosaurus is a Python library containing tools for automatic neural phonetic transcriptions using IPA (Li et al., 2020). Released by a team of researchers at Carnegie Mellon University, it performs phone recognition, rather than the more typical task of phoneme recognition, and its main innovation is its so-called "allophone layer". This is a phone-to-phoneme mapping that restricts the phonetic invntory to that of a single language. This enable Allosaurus to capture the benefits both of shared phoneme models and private phoneme models. While all three architectures use an encoder, shared phoneme models use a common phone layer across languages, while in the private phoneme model, each language has a separate model following the shared encoder. Allosaurus strikes a balance by using a shared decoder or "universal phone layer" and leaving the per-language customization to the allophone layer. This allows the decoder to benefit from larger amounts of training data and greater generality. These differences in architecture are shown below:



Graves et al. (2006)

3 Background

3.1 Automatic Speech Recognition

3.2 Text-to-Speech Synthesis

As its name suggests, test-to-speech synthesis involves generating waveform speech from text input. State-of-the art approaches are typically neural and involve a feature prediction network (FPN) and and audio generating network (AGN).

*** STILL NEED TO TYPESET REST OF NOTES ***

Székely et al. (2019)

3.3 Speech Signal Processing

Time-domain representations of signals are not ideal for most speech processing algorithms. To render speech more tractable, it is converted to the frequency domain, which allows a more compact representation, and one containing more information that is directly pertinent to the task of speech recognition.

3.3.1 Short-Time Fourier Transform

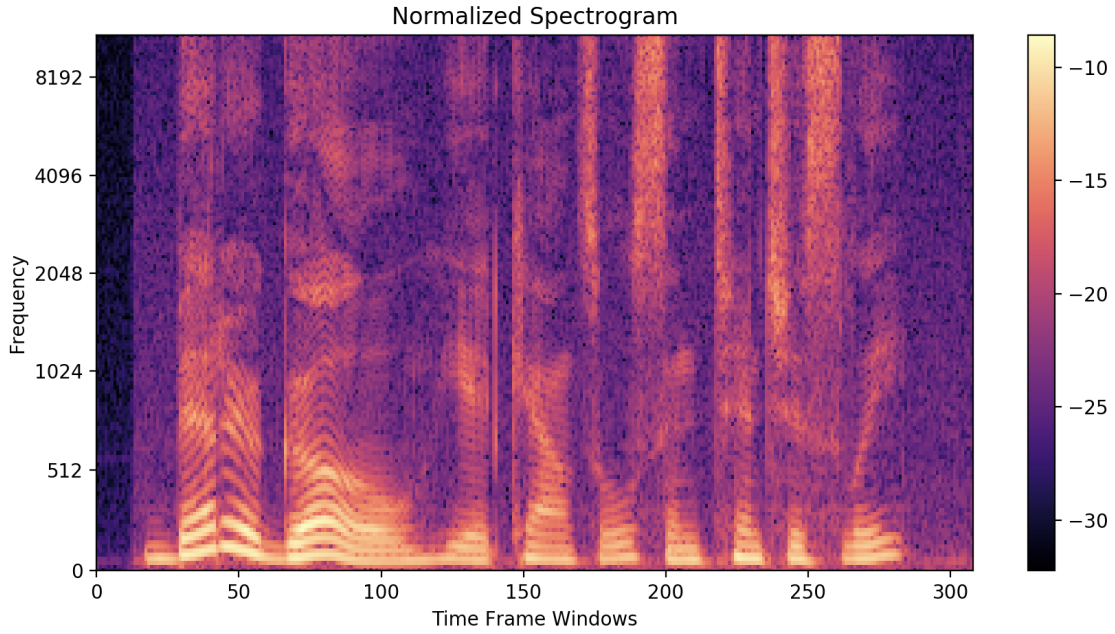
The favored approach to conversion from the tme domain to the frequency domain is the discrete-time short-time Fourier transform (STFT), which is given by

$$\mathbf{STFT}\{x[n]\}(m,\omega) := \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega m}$$

for some signal $x[n]$, windowing function $w[n-m]$, and frequency ω.

To avoid discontinuities between frames, i.e. between adjacent spectra, the window function advances across the signal at some fixed shift, which is less than the width of the window; in other words, the windows overlap.

To obtain the (power) spectrogram, the squared magnitude of each element is taken and the resulting discretized spectra are concatenated to form a matrix. Because this operation discards phase information, it is not lossless. As a result, a major task of speech synthesis from spectrograms is to reconstruct the phase. A sample spectrogram is shown below:



3.3.2 Mel Spectrum and Coefficients

A well-established finding from psychoacoustics is that human perception of sounds is not linear in frequency. Differences at lower frequencies are perceived as sounding 'further apart' than the same absolute difference at higher frequencies. To reflect this, speech is mapped nonlinearly onto the mel scale

$$M(f) = 1125 \ln(1 + f/700)$$

typically using overlapping triangular windows whose spacing reflects human perception.

To obtain the mel-frequency cepstrum, a number of additional operations are performed (Sahidullah and Saha, 2012). At each of the mel frequencies, the log is

taken of the power, which reflects the non-linear human perception of loudness. Finally, the Discrete Cosine Transform of these values gives the mel-frequency cepstral coefficients, which together make up the mel frequency cepstrum.

3.4 Neural Networks

Neural networks are, most generally, a powerful application of the ensemble approach to machine learning in which simple units ("nodes") are joined together to learn complex functions (Aggarwal et al., 2018). Originally inspired by the structure of the human brain, the degree to which this comparison is warranted remains controversial; nevertheless, the terminology is still used today.

At its core, a neural network consists of a series of linear and nonlinear operations that may be combined in various ways. Like other machine learning algorithms, an input is mapped to an output. However, a key advantage of neural networks is their ability to learn features internally.

Training is performed by means of the backpropagation algorithm, which makes possible the correct attribution of error to each parameter. The parameters are updated proportionally to their contribution to the error.

The simplest kind of neural network, the feedforward neural network (FFNN), consists of an *input layer*, zero or more *hidden layers*, and an *output layer*, where each hidden layer and optionally the output layer is passed to a nonlinear activation function such as $\sigma(x) = \dfrac{e^x}{1+e^x}$ or $\tanh(x) = \dfrac{e^x - e^{-x}}{e^x + e^{-x}}$. The activation function makes it possible to learn complex functions, since without it a feedforward neural network would always be equivalent to multiplication by a matrix.

3.4.1 Recurrent Neural Networks

Feedforward neural networks are effective in processing vector inputs. If the data possess a sequential nature, a FFNN may struggle to learn the intertemporal relationships. To solve this problem, the recurrent neural network was proposed, in which

at each time step, the network's output from that time step is passed as input to the network at the current time step. Notationally, given an input vector x at time t,

$$h(x_t) = f(h_{t-1}, x_t; \theta)$$

$$a_t = W^{(h \to h)} t_{t-1} + W^{(x \to h)} x_t + b$$

$$h_t = \tanh(a_t)$$

$$o_t = W^{(h \to o)} h_t + c$$

where f is the function computed by the RNN (Goodfellow et al., 2016). It is important that the loop inherent in the architecture of an RNN enables it to handle variable-length input well, which is an invaluable property in language- and speech-based settings where input and output length tend to vary.

One powerful and popular architecture for recurrent neural networks is the long short-term memory (LSTM) network (Hochreiter and Schmidhuber, 1997). One of the problems with simple RNNs is the difficulty of modeling long-term dependencies, and more generally with learning which information should be passed on ('remembered') and which information should be 'forgotten'. To remedy this, the LSTM adds a forget gate to regulate which information is passed on in the hidden state.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_t)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh(C_t)$$

The notation followed here is based on that of Olah (2015) and Staudemeyer and Morris (2019)

3.4.2 Sequence-to Sequence Architectures

A single RNN is well-suited to learning to classify an entire sequence or each time step of a sequence; in these respective settings, output size is fixed or equal to the input size. However, there are often settings in which input size and output size may differ and each varies. In such cases, a sequence-to-sequence model is used, in which two RNNs, often two LSTMs, are used. (Goodfellow et al., 2016)

3.4.3 Attention Mechanism

While LSTMs represent a considerable improvement over vanilla RNNs for a variety of tasks, they are far from perfect. One difficulty is the inherent difficulty of learning what to remember and what to forget. State information must be contained in a fixed-length vector. The attention mechanism circumvents this bottleneck by adding a sort of query mechanism by which the decoder has access to all hidden states of the encoder and learns weights to create a weighted sum, which is then used by the decoder to generate the output. This access to the entire input sequence allows the decoder to focus its 'attention', as it were, on the most relevant portions of the input, hence the name. This also brings the added benefit that the alignment between input and output can be examined and visualized.

3.4.4 Convolutional Neural Networks

When input data have a spatial component, it is beneficial to use a network architecture that can take advantage of spatial relationships in the data. A convolutional neural network does this by passing a $k \times l$-dimensional filter over the input. This is equivalent to a fully-connected network in which a majority of weights are zero. The relative sparsity of convolutional network networks (CNNs) is one advantage; the primary benefit is the ability, given multiple layers, to learn features that correspond to image features at various levels of resolution.

3.4.5 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 continually improves to generate increasingly realistic outputs, and the discriminator becomes increasingly adept at distinguishing real samples from generated samples.

3.5 Phonetics and the International Phonetic Alphabet

In natural human language, the correspondence between orthography and pronunciation is not entirely one-to-one; in some languages, such as English or Tibetan, it is difficult to predict pronunciation from spelling and vice versa. In some languages, such as Spanish or Swahili, there is a significantly higher degree of predictability, but the dialectical variation means that this tight correspondence does not hold for all groups of speakers. Additionally, even among languages using the same scripts, the same character often represents differents sounds in different languages. However, even within a language, characters can represent different speech sounds depending on surrounding letters or position within a word.

For all of these reasons, it is important to have an agreed-upon convention for representing speech sounds. The de facto standard is the International Phonetic Alphabet (Association et al., 1999; Hardcastle et al., 2012), which was developed to provide a language-agnostic (as well as dialect-agnostic) means for phonetic transcription. This makes it an excellent tool for transcribing accents in which differences in pronunciation are of interest.

The IPA is conceived to be able to represent, in theory, all phones appearing in any known language, or indeed, any phone realizable by humans. In the present work, because the language under consideration is English (albeit with two starkly differing accents), the full power of the IPA is not required. All segments considered are either vowels or consonants, and all consonants are pulmonic, i.e. produced by exhaling air from the lungs.

Vowels are classified along three key dimensions: openness/closeness, backness/frontness, and roundedness. The latter is binary, while the other two are continuous. Consonants are classified by place of articulation, manner of articulation, and voicedness. For example, b is the voiced bilabial plosive, while p is its unvoiced counterpart. Realized at a different place of articulation, g and k are the velar plosives (also respectively voiced and unvoiced). Alternatively, holding the place of articulation constant and changing the manner of articulation to fricative yields the pair β, ɸ. All phonetic segments used in this work fit this schema.

4 Methods

4.1 Phonetic Continuous Representation-Based Speech Synthesis and Accent Transfer [3]

4.1.1 Phonetic Posteriorgrams

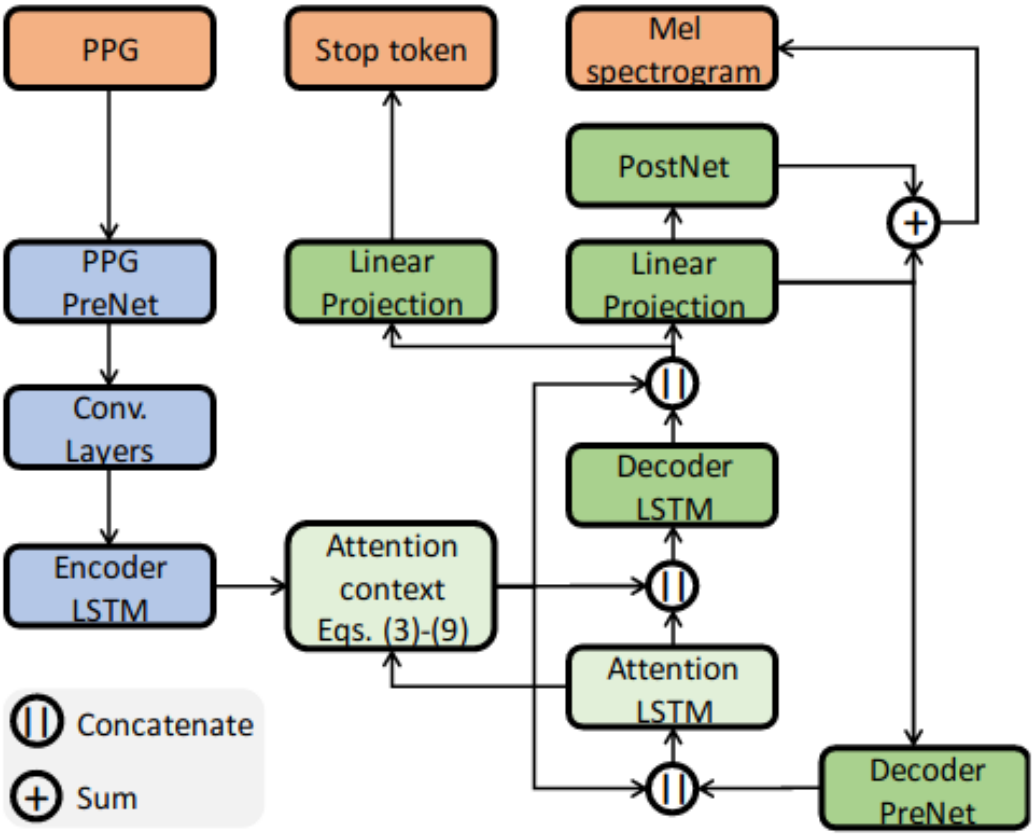
A phonetic posteriorgram represents predicted framewise phoneme probabilities from a speech sample. In this work, a pretrained acoustic model from [XX] was used, due to the high quality of the model, which had been trained on the extremely large and varied LibriSpeech corpus.

The acoustic model predicts 5816 senones; the dimensionality is greatly reduced by a linear discriminant analysis matrix, yielding the posterior probability for the 40 phonemes in the TIMIT reduced phoneme set, which include a silence phoneme (not a true phoneme, but essentially treated as one). Refer to the Data section for samples.

4.1.2 Speech Synthesis from PPGs [1]

To generate a mel spectrogram representation from a PPG, a modified version of Tacotron2 was used. This worked followed NVIDIA's open-source implementation with a few modifications. As in [XX], and as shown below, the key differences were the addition of a localized attention constraint, which decreases the computational complexity. This is especially attractive in this case because the expected alignment between PPG and mel spectogram is tighter than the alignment between text and mel spectrogram (e.g. in the standard Tacotron2 model). Because of this, it is superfluous to query all encoder states.

The other important difference is that instead of an embedding layer for the input, a linear projection is used, since the input is not sparse, i.e. not one-hot encoded. Thus, each input dimension must be taken into account by the model. The following image summarizes the architecture [XX]:



4.1.3 Accent Transfer in PPG Space [2]

Inspired by the basic architecture of MelGAN-VC, the PPG accent transfer model uses a generative adversarial network with a cyclical consistency constraint. There are two different variants.

The first architecture is convolutional, using one-dimensional convolutions across time. Two-dimensional convolutions were not used because there are no local spatial relationships across phonemes in a PPG; phoneme oder is not meaningful. The spatial relationships of interest are exclusively temporal.

The second architecture is sequential, a bidirectional LSTM. Each of these architectures, when used, is used in both the discriminator and the generator.

*** ILLUSTRATION TO COME ***

4.2 Phonetic Transcription-Based Speech Synthesis and Accent Transfer

4.2.1 Relational Mappings for Segmental Accent Transfer [1]

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.

A phonetic sequence consists of a number of IPA segments s_i:

$$S^{(n)} = \langle \text{start} \rangle \, , s_1^{(n)}, s_2^{(n)}, ...s_L^{(n)}, \langle \text{stop} \rangle$$

Given the data, it is straightforward to calculate the empirical conditional probabilities. The probability that segment s_j follows segment s_i is calculated as follows:

$$\hat{p}(s_j|s_i) = \frac{c_{ij} + 1}{c_i}$$

where c_{ij} represents the corpus frequency of the bigram s_i s_j and c_i represents the corpus frequency of s_i. This is analogous to a simple language model with smoothing, but it would be more accurate to refer to the full set of conditional probabilities as a phonetic model.

Another necessary component is a matrix of segment similarities. This can be written as

$$A = [a_{ij}]$$

where

$$a_{ij} = \mathrm{sim}(s_i, s_j) = \frac{\mathbf{f}_i \cdot \mathbf{f}_j + 1}{||\mathbf{f}_i||\ ||\mathbf{f}_j||},$$

i.e. each element of this matrix represent the smoothed cosine similarity product of the appropriate articulatory feature vectors. For the special characters _ (word boundary), ⟨start⟩, and ⟨stop⟩, define self-similarity $\mathrm{sim}(x, x) = 1000$, since each of these characters will invariably be mapped to itself.

The following table lists and describes the articulatory features considered.

Articulatory Features

Feature	Relevant Segments
voiced	a, b, etc.
voiceless	f, h, etc.
vowel	a, æ, etc.
approximant	l, l̪, ɹ, ɻ, w
consonant	b, d, ð, etc.
front vowel	a, æ
central vowel	ə, ɨ
back vowel	ɑ, ɒ, o, ɔ, u, ʊ, ʌ
open vowel	a, æ, ɑ, ɒ, ɛ, ɔ, ʌ
mid vowel	e, ə, ɛ, o, ɔ, ʌ
close vowel	e, i, ɪ, ɨ, o, u, ʊ
unrounded vowel	a, æ, ɑ, e, ɛ, i, ɪ, ɨ, ʌ
rounded vowel	ɒ, o, ɔ, u, ʊ
bilabial	b, m, p, w
labiodental	f, v, w
dental	ð, l̪, θ
alveolar	d, ʤ, l, n, ɹ, ɻ, ɾ, s, ʃ, t, tʰ, ʧ
postalveolar	ʃ, ʒ
retroflex	ɖ, ɳ, ʂ, ʈ, ʐ
palatal	j, ʝ, ɟ
velar	g, k, kʰ, ŋ, x
uvular	none
pharyngeal	none
glottal	h

Articulatory Features (cont.)

Feature	Relevant Segments
plosive	b, d, ɖ, ɟ, k, k^h, p, t, t^h, ʈ
nasal	m, n, ɳ, ŋ
trill	r
tap/flap	ɾ
fricative	ð, f, h, j, s, ʂ, ʃ, v, x, z, ʐ, ʒ
lateral	l, ɭ
affricate	ʤ, ʧ
rhotic	r, ɹ, ɻ, ɾ

The core problem of segmental accent transfer is to learn a set of segmental mapping rules. In the simplest case, such a mapping is injective.When an injective mapping is assumed, the mapping can be defined as follows:

$$f_1(s_i^{(n)}) = t_i^{(n)} = \arg\max_{t_j} \alpha \cdot p(t_j|t_{j-1}) + \text{sim}(t_j, s_i^{(n)}).$$

Here α controls the weighting of the target phonetic model relative to similarity between source and target segments. It is to be chosen empirically.

Equivalently, in pseudocode notation:

Algorithm 1 Injective Mapping

input: list S, function sim(), phone model p, target phone inventory I, $\alpha \in [0, 1]$
output: list T
$i \leftarrow 1$
$T[0] \leftarrow \langle \text{start} \rangle$
$N \leftarrow \text{length}(S)$
while $i < N$ **do**
    best $= 0$
    for phone in I **do**
        score $= \alpha * p(\text{phone} \,|\, T[i-1]) + \text{sim}(\text{phone}|S[i])$
        if score $>$ best **then**
            best $\leftarrow$ score
            phone* $\leftarrow$ phone
        end if
    end for
    $T[i] \leftarrow$ phone*
    $i \leftarrow i + 1$
end while
$T[i] \leftarrow \langle \text{stop} \rangle$

Note that this belongs to the family of greedy search algorithms.

However, it may be desirable to increase flexibility by allowing one-to-two and two-to-one mappings. To accomplish this, it is necessary to add to the target character set the empty string. As in the formal language theory literature, the character "ϵ" will denote the empty string. To allow 2-to-1 mappings, define

$$\text{sim}(\epsilon, s) := \bar{a} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij}}{N^2} \quad \forall s \in S : s \neq \epsilon$$

and

$$\text{sim}(\epsilon, \epsilon) = 1$$

To allow for more general alignments, we define a more flexible algorithm as follows:

Algorithm 2 Bigram-Compatible Mapping

input: list S, function sim(), phone model p, target phone inventory I, $\alpha \in [0, 1]$
output: list T
$i \leftarrow 1$
$T[0] \leftarrow \langle \text{start} \rangle$
$N \leftarrow \text{length}(S)$
while $i < N$ **do**
    best $= 0$
    for phone in I **do**
        score1 $= \alpha * p(\text{phone} \,|\, T[i-1]) + \text{sim}(\text{phone}|S[i])$
        if score1 $>$ best **then**
            best $\leftarrow$ score1
            phone* $\leftarrow$ phone
            $i \leftarrow i + 1$
            outcome $\leftarrow 1$
        end if
        score2 $= \alpha * p(\text{phone} \,|\, T[i-1]) + \text{sim}(\text{phone}|S[i-1])$
        if score2 $>$ best **then**
            best $\leftarrow$ score2
            phone* $\leftarrow$ phone
            outcome $\leftarrow 2$
        end if
        score3 $= \alpha * p(\text{phone} \,|\, T[i-1]) + \text{sim}(\text{phone}|S[i+1])$
        if score3 $>$ best **then**
            best $\leftarrow$ score3
            phone* $\leftarrow$ phone
            outcome $\leftarrow 3$
        end if
    end for
    if outcome < 3 **then**
        $T[j] \leftarrow$ phone*
        $j \leftarrow j + 1$
    end if
end while
$T[i] \leftarrow \langle \text{stop} \rangle$

Note that the modifications relative to Algorithm 1 allow for both 1:2 and 2:1 mappings. For example, if in the relation R it is the case that $s_{i-1}rt_{j-1}$, it is possible at alignment step (i, j) to have

$a)$ $s_{i-1}Rt_{j-1}, s_iRt_j$ (outcome1)
$b)$ $s_{i-1}Rt_{j-1}, s_{i-1}Rt_j$ (outcome2)
$c)$ $s_{i-1}Rt_{j-1}, s_iRt_{j-1}$ (outcome3)

Thus, it is possible to model systematic insertions and deletions, without increasing computational complexity, which remains $\mathcal{O}(n)$.

The algorithm can additionally be enhanced by

Finally,

4.2.2 Segment-Based Speech Synthesis

Once again, a modified version of Tacotron2 is used to generate mel spectrograms, and Waveglow is used for waveform generation. The input encoding is different from the classic English speech synthesis implementation, but otherwise no modifications were required.

5 Data

5.1 Speech Corpora

To apply these methods, a novel dataset was created, containing speech recordings of two speakers. All data were downloaded from YouTube using the open-source tool youtube-dl. The first speaker is Grant Sanderson, a well-known educator and popularizer of mathematics. His dialect is standard American English. The second is Rajamunickam Antonimuthu, a YouTuber operating a popular channel about science and technology news. His native language is Tamil, and is has a srong influence on his speech. The speech samples of each speaker are relatively clean spontaneous speech. Depending on the exact process each uses for video creation, it may be at least

partially scripted, but the nature of the speech samples is closest to spontaneous speech, especially compared to professional-quality read speech such as the LJSpeech corpus. In particular, there are more pauses and filler words.

For recognition and especially for synthesis, it is crucial to divide the longer speech files into smaller segments. To do this for the Antonimuthu corpus, subtitles were also scraped from the videos. Because the subtitles are time-aligned to the video, and hence to the audio files, it was possible to divide on subtitle time stamps, combining segments up to a maximum length of 10 seconds.

For the Sanderson corpus, no subtitles were available, and so it was necessary to cut on silence. To accomplish this, the Python library `pydub` was used, with a silence threshold of -30 and a minimum silence length of 300 miliseconds. These values were found empirically to work best. As in the Atonimuthu corpus, smaller segments were combined as long as their combined length was less than 10 seconds.

The final corpus sizes were 10500 utterances for Sanderson and 2100 utterances for Antonimuthu, with average utterance length 4.49 and 6.34 seconds, respectively.

5.2 Mel Spectrograms

The mel spectrograms were generated in PyTorch as in the original Tacotron2 and Waveglow open-source implementations by NVIDIA [XX]. The hyperparameters were as follows:

- input sampling rate: 22050 Hz
- number of mels: 80
- window length: 1024 (samples) $\approx$ 46.44ms
- filter length: 1024 (samples) $\approx$ 46.44ms
- shift (hop length): 160 (samples) $\approx$ 7.26ms
- minimum mel frequency: 0 Hz

- maximum mel frequency: 8000 Hz
- audio bit depth: 32768 (15-bit)

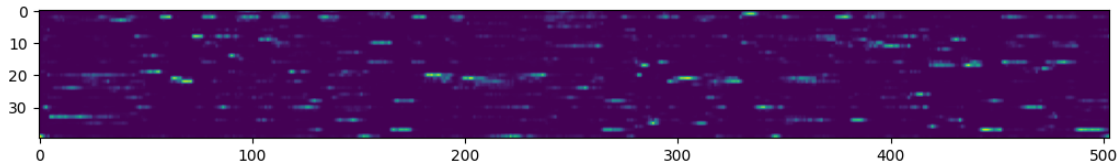
5.3 Phonetic Posteriorgrams

Using the acoustic model described earlier and introduced in [XX], phonetic posteriorgrams were generated for each utterance, using the following hyperparameters:

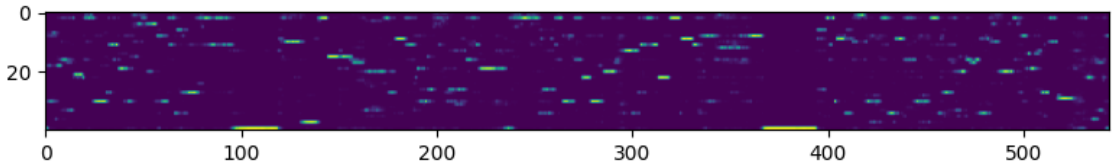
- input parameters:
  - number of mels (for input): 80
  - sampling rate: 22050 Hz
  - window length: 64ms
  - filter length: 64ms
  - shift (hop length): 10ms
- number of senones: 5816
- number of output phonemes: 40
- phoneme set: TIMIT reduced (40-phoneme) set

Examples:

- An Antonimuthu PPG representing the utterance *"To overcome the drawbacks of SLM, Lawrence Livermore National University researchers"*:



- A Sanderson PPG representing the utterance *"Geometric algebra, that is a whole other topic, uh, potentially for another day, but that's a - that's a really valid qustion"*:



5.4 Transcriptions [1]

The transcription process made use of a hybrid process. Analogously to the gain in speed reported in machine-assisted translation, it is faster to correct automatically-generated transcriptions than it is to manually create all transcriptions from scratch.

First, transcriptions were generated automatically by the Python library Allosaurus. These were only moderately accurate, and due to the nature of the Connectionist Temporal Classification loss which which the model is trained, the output sequence cannot have repeating tokens. Additionally, word boundaries were not included. This lack of word boundaries was hypothesized to be disadvantageous for speech synthesis, especially because the standard Tacotron 2 model input includes spaces and (standardized) punctuation.

Because of this, transcriptions were corrected by hand. Word boundaries were added, denoted by the underscore character (_). Additionally, the removal of repeated phones in different words was corrected. For example, the phrase "*that takes some more rice*" would be transcribed as [ð æ t e k s ʌ m o ɹ a ɪ s]. Following the manual correction, the transcription would be [ð æ t _ t e k s _ s ʌ m _ m o ɹ _ ɹ a ɪ s]. Intuitively, it appears clear that the second transcription contains more information that would be relevant for IPA-to-speech synthesis.

Because successful automatic synthesis requires many examples of each input token, for each speaker the rarest IPA segments were removed and replaced with the best more common alternative, leaving only those segments occurring at least 20 times in the fine-tuning set of 200 utterances.

To improve the automatically-generated transcriptions, the pretrained models released by the authors of XX were fine-tuned on 200 manually corrected transcriptions each. Subsequently, an additional XX utterances were transcribed for each speaker, creating two datasets with XX utterances each.

Phonetic Inventories

Phone	Definition	Antonimuthu	Sanderson
a	open front unrounded vowel	✓	✓
æ	near-open front unrounded vowel	✓	✓
ɑ	open back unrounded vowel	✓	✓
ɒ	open back rounded vowel	✓	
b	voiced bilabial plosive	✓	✓
d	voiced alveolar plosive	✓	✓
ð	voiced dental fricative	✓	✓
ʤ	voiced post-alveolar affricate	✓	✓
ɖ	voiced retroflex plosive	✓	
e	close-mid front unrounded vowel	✓	✓
ə	mid central vowel (schwa)	✓	✓
ɛ	open-mid front unrounded vowel	✓	✓
f	voiceless labiodental fricative	✓	✓
g	voiced velar stop	✓	✓
h	voiceless glottal fricative	✓	✓
i	close front unrounded vowel	✓	✓
ɪ	near-close front unrounded vowel	✓	✓
ɨ	close central unrounded vowel	✓	
j	voiced palatal approximant (yod)	✓	✓
ʝ	voiced palatal fricative	✓	
ɟ	voiced palatal plosive	✓	
k	voiceless velar plosive	✓	✓
k^h	aspirated voiceless velar plosive	✓	
l	voiced alveolar lateral approximant	✓	✓
l̪	voiced dental approximant	✓	
m	voiced bilabial nasal	✓	✓

Phonetic Inventories (cont.)

Phone	Definition	Antonimuthu	Sanderson
n	voiced alveolar nasal	✓	✓
ɳ	voiced retroflex nasal	✓	
ŋ	voiced velar nasal	✓	
o	close-mid back rounded vowel	✓	✓
ɔ	open-mid back rounded vowel	✓	✓
p	voiceless bilabial plosive	✓	✓
r	voiced alveolar trill	✓	
ɹ	voiced alveolar approximant		✓
ɻ̩	syllabic voiced alveolar approximant	✓	
ɾ	voiced alveolar tap/flap	✓	
s	voiceless alveolar fricative	✓	✓
ʂ	voiceless retroflex fricative	✓	
ʃ	voiceless postalveolar fricative	✓	✓
t	voiceless alveolar plosive	✓	✓
tʰ	aspirated voiceless alveolar plosive	✓	
ʧ	voiceless palato-alveolar affricate	✓	✓
ʈ	voiceless retroflex plosive	✓	
u	close back rounded vowel	✓	✓
ʊ	near-close back rounded vowel	✓	✓
v	voiced labiodental fricative	✓	✓
ʌ	open-mid back unrounded vowel	✓	✓
w	voiced labial-velar approximant	✓	✓
x	voiceless velar fricative	✓	
z	voiced alveolar fricative	✓	✓
ʐ	voiced retroflex sibilant fricative	✓	✓
ʒ	voiced postalveolar fricative		✓
θ	voiceless dental (non-sibilant) fricative		✓

6 Results

6.1 Continuous Representation-Based Speech Synthesis and Accent Transfer

The usefulness of pre-synthesis transformation of PPGs requires that there be a systematic difference between the PPGs of speaker A and speaker B. Empirically, this was not the case, as the loss of each classifier converged around 0.69 with only small and apparently random fluctuations even after hundreds of epochs. This value is approximately the expected loss of binary cross entropy on random data, or $-\ln\frac{1}{2} \approx 0.693147$. Thus, we can confidently conclude that there is no systematic difference between the PPGs of Sanderson and Antonimuthu.

While disappointing, this finding was not entirely uninteresting. It suggests that, at least at the resolution of 40 phonemes, even speakers speaking the same language with starkly contrasting accents will have their speech mapped into a common phonetic space, in which the respective phonetic representations are indistinguishable. This negates one of the core initial hypotheses of this work, which was that for strongly differing accents, an acoustic model will

A further disappointing finding was that speech synthesis from 40-phone PPGs was not able to generate coherent speech for the Sanderson and Antonimuthu corpora. Because Waveglow generated comprehensible speech from mel spectrograms, this was clearly due to the generation of mel spectrograms from PPGs. This falsifies the original hypothesis that high-quality speech synthesis from 40-phone PPGs is possible, at least from spontaneous speech training data.

While the "black box" nature of neural networks makes it impossible at this point to draw conclusions about the causes of this failure with any certainty, there are two factors that seem most likely. First, it may be that the 40-phone PPGs simply do not carry sufficient information. As described above [XXX] succeeded in generating high-quality speech from 5816-dimensional PPGs. The reduction by more than two orders of magnitude may simply too large a reduction.

Secondly, the nature of the data in the respective speech corpora is, despite the data cleaning process, noisier and less uniform than standard corpora used in speech synthesis, such as LJSpeech. This is primarily due to the spontaneous nature of the speech, which is is more difficult to re-create with a neural network than speech read by a professional voice actor.

6.2 Transcription-Based Speech Synthesis and Accent Transfer

7 Conclusions

It should be noted that this work does not make any normative prescriptions about a "correct" dialect of English, nor does it condone such prescriptions. In the author's view, different dialects are equally valid. Applications of accent transfer, such as accent reduction, may be motivated by practical concerns, such as the desire to decrease the difficulty of understanding or being understood by an interlocutor. However, these considerations do not imply the superiority of any accent over another. The focus on one of the two theoretically possible directions of transfer, namely Indian English to American English, was motivated by considerations regarding the phones available in each.

This work, unfortunately, was not able to improve upon the results of Zhao et al. (2019) regarding speech synthesized from PPGs. However, this work does suggest a number of directions for future research. an interesting direction for future work might be to work with multi-accent or even multi-language acoustic models. Conceivably, speakers with different accents would be distinguishable in such a phonetic space, rendering the transfer models more useful.

The failure of the modified Tacotron models to synthesize coherent speech is also in need of further investigation. One recommended line of inquiry, which this work would have explored if time had permitted, regards whether prompted speech by a professional voice actor would yield better results.

The two most interesting novel contributions in this work are the demonstration of speech synthesis from IPA, and the presentation of algorithms for transcription-to-transcription mappings. Each of these also points to the potential for interesting work to be done. Neural speech synthesis from IPA, while obviously more challenging than established methods, offers exciting possibilities. It would represent a further step toward the ideal of a multilingual, multi-speaker speech synthesis system, which in turn would play an important part in end-to-end spech translation. It also has the potential to be used in smaller practical applications, such as accent transfer and human accent training.

*** TO BE ADDED - WAITING FOR RESULTS ***

References

Charu C Aggarwal et al. Neural networks and deep learning. *Springer*, 10:978–3, 2018.

Sandesh Aryal and Ricardo Gutierrez-Osuna. Accent conversion through cross-speaker articulatory synthesis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7694–7698. IEEE, 2014.

Sandesh Aryal and Ricardo Gutierrez-Osuna. Comparing articulatory and acoustic strategies for reducing non-native accents. In *INTERSPEECH*, pages 312–316, 2016.

Sandesh Aryal, Daniel Felps, and Ricardo Gutierrez-Osuna. Foreign accent conversion through voice morphing. In *Interspeech*, pages 3077–3081, 2013.

International Phonetic Association, International Phonetic Association Staff, et al. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

Fadi Biadsy, Ron J Weiss, Pedro J Moreno, Dimitri Kanevsky, and Ye Jia. Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation. *arXiv preprint arXiv:1904.04169*, 2019.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

William J Hardcastle, John Laver, and Fiona E Gibbon. *The handbook of phonetic sciences*, volume 119. John Wiley & Sons, 2012.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Takuhiro Kaneko and Hirokazu Kameoka. Parallel-data-free voice conversion using cycle-consistent adversarial networks. *arXiv preprint arXiv:1711.11293*, 2017.

Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions, 2018.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, et al. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253. IEEE, 2020.

Christopher Olah. Understanding lstm networks. 2015.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621. IEEE, 2019.

Md Sahidullah and Goutam Saha. Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. *Speech communication*, 54(4):543–565, 2012.

Ralf C. Staudemeyer and Eric Rothstein Morris. Understanding lstm – a tutorial into long short-term memory recurrent neural networks, 2019.

Éva Székely, Gustav Eje Henter, Jonas Beskow, and Joakim Gustafson. Spontaneous conversational speech synthesis from found data. In *INTERSPEECH*, pages 4435–4439, 2019.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio, 2016.

Guanlong Zhao and Ricardo Gutierrez-Osuna. Using phonetic posteriorgram based frame pairing for segmental accent conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(10):1649–1660, 2019.

Guanlong Zhao, Sinem Sonsaat, John Levis, Evgeny Chukharev-Hudilainen, and Ricardo Gutierrez-Osuna. Accent conversion using phonetic posteriorgrams. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5314–5318. IEEE, 2018.

Guanlong Zhao, Shaojin Ding, and Ricardo Gutierrez-Osuna. Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams. In *Proc. Interspeech 2019*, pages 2843–2847, 2019. doi: 10.21437/Interspeech.2019-1778. URL `http://dx.doi.org/10.21437/Interspeech.2019-1778`.