A neural network-based approach to accent conversion

Kenny W. Lino

M.Sc. Dissertation

Department of Intelligent Computer Systems
Faculty of Information and Communication Technology
University of Malta
2018

Supervisors:
Claudia Borg, Department of Artificial Intelligence, University of Malta
Andrea DeMarco, Institute of Space Sciences and Astronomy, University of Malta
Eva Navas, Department of Communications Engineering, University of the Basque Country

Submitted in partial fulfilment of the requirements for the Degree of
European Master of Science in Human Language Science and Technology

M.Sc (HLST)

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY UNIVERSITY OF MALTA

Declaration

Plagiarism is defined as ”the unacknowledged use, as one’s own work, of work of another person, whether or not such work has been published” (Regulations Governing Conduct at Examinations, 1997, Regulation 1 (viii), University of Malta).

I, the undersigned, declare that the Master’s dissertation submitted is my own work, except where acknowledged and referenced.

I understand that the penalties for making a false declaration may include, but are not limited to, loss of marks; cancellation of examination results; enforced suspension of studies; or expulsion from the degree programme.

Student Name: Kenny W. Lino
Course Code: CSA5310 HLST Dissertation
Title of work: A neural network-based approach to accent conversion

Signature of Student:

Date:

Abstract

With the emergence of the use of technology in language learning through tools like Rosetta Stone and Duolingo, learners have slowly been given more autonomy of their language learning projection. Although these tools have allowed learners to tailor their learning to their own liking, there is a gap between the available resources to assist those that would like to improve their pronunciation. Previous research in the intersection of language learning and speech technology has made efforts to develop pronunciation training systems to address this problem, but the systems themselves tend to have gaps due to the lack of appropriate support for the users, especially in appropriately identifying errors and providing sufficient feedback to help them correct their errors.

Some researchers have purported that alongside other forms of feedback such as a visual articulatory representation, a voice conversion system could serve as a potential feedback mechanism by helping learners understand what their voice could sound like given the appropriate changes. However, like pronunciation training systems, voice conversion systems also faced many limitations especially in terms of the quality which made them unrenderable as useful tools. With that said, recent advances in speech technology using deep neural networks have become increasingly successful in achieving better accuracy and quality in a variety of tasks, allowing for the potential to return and address these said gaps in quality and performance for voice conversion.

This dissertation investigates these advancements in applying deep neural networks to develop a voice conversion system that could potentially serve as a feedback mechanism as a part of a larger computer-based pronunciation training system. Specifically, I intend to adapt the methodologies of Aryal and Gutizerrez-Osuna (2014) to set forth an accent conversion system that strives to convert a source voice into a target accent, leveraging neural network architectures in place of Gaussian Mixture Models for conversion.

Contents

List of Figures

List of Abbreviations

CAPT	Computer Assisted Pronunciation Training
CP	Critical Period
L1/L2	First and second language
LSTM	Long-short term memory
MFCC	Mel-frequency cepstrum coefficient TTS
Text-to-speech	

Chapter 1

Introduction

Technology has continuously evolved to no bounds as witnessed by the current successes enjoyed by the use of neural networks and the power of current hardware, something perhaps predicted by Moore's Law who proclaimed that computing power would double once every 18 months (and then changed to 24 months) [CITE HERE]. [Mention something about AI here]

We see the effects of neural networks throughout many subareas in computer science, including that of natural language processing. In fact, if we take a look at the number of publications involving neural networks, it has exponentially compounded annually [CITE IMAGE HERE].

While technology has flourished and led to a number of new state-of-the-art systems such as improvements in commercial speech recognition and machine translation, it can be argued that these benefits have not reached and innovated other areas outside of research to the same extent. One such example that could benefit from modern innovations is education. Although there have been small trends here and there to create applications for educational use such as Duolingo for language learning[EXAMPLES?], in general it seems that education has not evolved at the same rate as technology. In particular, pronunciation has been a large standing challenge in language learning due to its complex nature. Unlike grammar and vocabulary, pronunciation can be challenging to both learn and teach due to the lack of clarity on how to teach it. Like grammar and vocabulary, pronunciation also involves a number of nuanced characteristics, including stress, rhythm, vowels, and consonants.

The variation in these features contribute to what many known as *accent*, or variations in pronunciation across speakers based on location, ethnicities, social classes, native languages, etc. Accents can be considered to be a part of dialects, where users of the same language may have variations beyond pronunciation, such as usage in vocabulary or grammar. The line may often be blurred in everyday discussions and even in academic analyses as accent and dialect (as well as language) could be considered to be on a continuum, but for the sake of simplicity, I consider *accent* to be variations in pronunciation in this work.

1.1 Research Questions

In this thesis, I focus on investigating the following questions:

- How can we leverage recent advances in recent technologies (namely deep neural networks) to convert a speaker's voice into sounding like it was said with another accent?
- What specific methodologies achieve the best similarity to the target accent and produce the most natural sounding audio?

1.2 Thesis Overview

The overview of the thesis is as follows:

In **Chapter 2**, I give a proper definition of voice conversion and accent conversion, and a high level overview of some technical details needed to better understand the current work.

In **Chapter 3**, I present the motivation for creating an accent converison system by discussing previous findings in second language acquisition research especially in relation to speech. I then cover previous work in voice and accent conversion to frame the advances and shortcomings of previously developed systems.

In **Chapter 4**, the design and methodology of the experiments are presented alongside the appropriate tools utilized to conduct each one.

In **Chapter 5**, the results of the experiments previously described are presented along with some short discussion and conclusions drawn from the results.

In **Chapter 6**, the thesis is concluded with a reflection on the work presented along with some appropriate suggestions for future work.

Chapter 2

Background

Before delving into previous literature and their relevance to this work and the fields of NLP and language learning as a whole, I detail both voice conversion and accent conversion in order to help better distinguish them. I also go over some common speech technology concepts typically used in these systems at a high level in order to make the current work more accessible to those unfamiliar with the area. Further reference is also provided for those interested in the technical aspects and formalisms.

2.1 Voice conversion

To properly frame voice conversion, we take a look at Mohammadi and Kain (2017) who present a recent overview of the subfield. Following a definition setforth by the authors, voice conversion refers to the transformation of a speech signal of a *source speaker* to make it sound similar to a *target speaker* in any chosen fashion with the utterance still being intact. Some of these changes can include changes in emotion, accent, or phonation (whispered/murmured speech). there have been a number of proposed uses for VC, including the transformation of speaker identity (perhaps for voice dubbing), personalized TTS systems, and against biometric voice authentication systems.

Voice conversion often involves a large number of processes, one of which includes deciding the appropriate type of data. To start, one must decide whether to have parallel or non-parallel speech data. Parallel speech data refers to speech data that has source and reference speakers that say the same utterance, so only the speaker-specific information is different, while non-parallel data would indicate datasets where the utterances are not the same, and thus entail further processes to create a target waveform. Even though parallel corpora are more desirable as it reduces the footprint necessary for conversion, parallel corpora are often curated for specific purposes and are not available in most cases.

Because of its simplicity, in some cases, researchers have tested making a psuedo-parallel corpus using acoustic clustering when working with non-parallel data (Lorenzo-Trueba et al. 2018; Sundermann et al. 2006)

Other aspects that need to be considered as discussed by Mohammadi and Kain (2017) include whether the data is *text-dependent* or *text-independent*. Text-dependent corpora indicate that the data has word or phonetic transcription, which can ease the alignment process during training, while systems using text-independent data would need to find similar speech segments, using algorithms such as clustering before training. Finally, one minor aspect that is not considered often is the languages of the source speaker and target speaker. Although many systems tend to focus on voice conversion between two native speakers of the same language, systems that aim to convert between two speakers speaking in different languages would have to be wary of potential mapping issues between sounds. This is especially important to consider in terms of accent conversion, which will be discussed in the next section.

Aside from considering these aspects of the corpora, the type of features extracted from the waveforms heavily impact the quality of the conversions. In investigating the most salient features of speaker individuality, previous researchers have concluded that the average spectrum, formants, and average pitch level are the most relevant. Following these conclusions, most VC systems focus on converting these features, and often work at the frame-level (windows of ˜20ms), with the assumption that the frame represents a stationary sound. From these frames, there a number of common *local* features that are extracted to represent the speech. These include the spectral envelope, cepstrum, line spectral frequencies (LSF) and the aforementioned formants. In particular, working with the mel-frequency cepstrum coefficents (MFCCs) are very standard in not just voice conversion systems, but most speech synthesis and recogntion systems in general. These are described in further detail in the section titled [FEATURES?; add label here to add easy click.]

On top of these local frame-based features, contextual features can be considered as well, although this would entail further fine-tuning of the features and system. [expand]

A visual representation that summarizes the voice conversion process can be seen in Figure 2.2, courtesy of Mohammadi and Kain (2017).

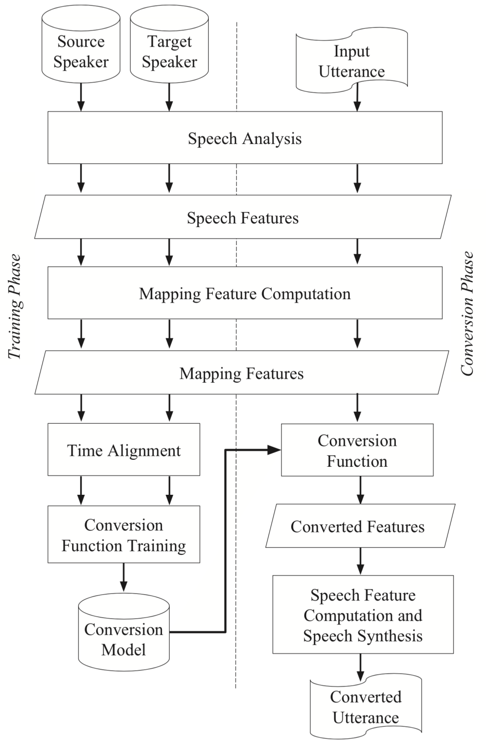

Figure 2.1: The training and conversion processes of a typical VC system.

2.2 Accent conversion

Like voice conversion, accent conversion is dedicated to convert the speech of a *target speaker* into sounding more like a *source speaker*. However, accent conversion is specifically focused on morphing the *accent* of the speech signal, as opposed to sounding directly like the source speaker. Succinctly stated, "Accent conversion seeks to transform second language L2 utterances to appear as if produced with a native (L1) accent," (Aryal and Gutierrez-Osuna 2014a). Accent conversion poses a further challenge on top of (parallel) voice conversion as the audio of the source speaker and target speaker is often forced-aligned. This means that with native and non-native speech, voice conversion would retain the voice quality and accent of the target speaker (Aryal and Gutierrez-Osuna 2014b).

2.3 Technical Background

2.3.1 Mel-frequency cepstrum coefficients

Following Jurafsky (2009), mel-frequency cepstrum coefficients (MFCCs) allow us to create vectorized representation of the acoustic information.

This is done by going over the speech signal using *windows*, where each window is assumed to contain a non-changing part of the signal. In order words, each window would roughly contain one phone– or speech sound. In order to retain all of the necessary information from each part of the signal, the windows often overlap.

After the signal is separated into different windows, the spectral information can be extracted using a special tool or formula known as the Discrete Fourier Transform. This allows us to find how much energy is in specific frequency bands.

From here the frequencies outputted by the Discrete Fourier Transform are converted onto the *mel* scale, which is where the *mel* in Mel-frequency comes from. In short, the mel scale is used to represent human hearing, which is more sensitive to lower pitch sounds (under 1000hz) as compared to higher pitch sounds. Afterwards, the *cepstrum* is calculated in order to separate source information from filter information. From a high level, the source-filter theory says that all sounds come from the glottis (the area around our throat) and below, which contains information common to all speech sounds, such as the fundamental frequency (or pitch) of someone's voice, as well as glottal pulse information. This is compared to the filter, which says that adjusting the vocal tract (e.g. moving the tongue and other articulators) define each individual sounds. By retaining just the filter information, we can model an individual phone.

A visual representation of the whole MFCC extraction process can be seen in Figure 2.2, taken from Jurafsky (2009).

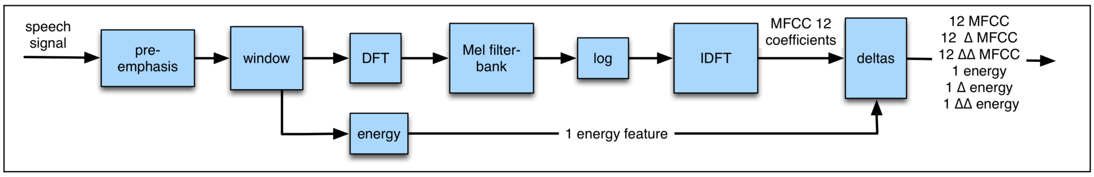

Figure 2.2: The extraction of sequence 39-dimensional MFCC vectors fro a waveform.

2.3.2 Gaussian mixture models

2.3.3 I-vectors

Chapter 3

Literature Review

This section provides a brief overview of second language acquisition and education in order to motivate the usage of technology in language learning. I then examine some previous research in computer assisted pronunciation (CAPT) systems in order to motivate discussion about voice conversion and accent conversion, where I detail important pivotal work done in the area.

3.1 Theoretical and educational motivations

Linguists have long debated over the possibility of whether second language (L2) learners (e.g. adult learners) could ever acquire a language to the extent of a native speaker. Some still cite ideas like the Critical Period (CP) Hypothesis and neuroplasticity which claims that learners cannot acquire language (at least as well as a native speaker) after a certain point in time due to the loss of plasticity in the brain (Lenneberg 1967; Scovel 1988). This theory has been particularly cited in reference to pronunciation, perhaps due to the obvious difficultly in overcoming the L1 negative transfer that many, if not all, language learners experience in speaking a new language.

Since the emergence of the CP hypothesis, many researchers have come to find evidence that suggest the contrary. In Lengeris (2012), we are presented an overview of the interactions between factors that affect second language acquisition such as age, linguistic experience, and learning setting. Here, we find evidence of studies such as Bongaerts et al. (1995), which present a counterargument against the CP hypothesis. In this study, they discovered through a foreign accent rating study with Dutch learners of English that learners could be perceived as *indistinguishable* from native speakers. Other researchers such as Flege have also found that there is no distinct 'cut-off' point like the CP suggests. Thus, while age may have some effect on a speaker's pronunciation, there is no conclusive evidence to say that the loss of plasticity in the brain leads to an inability to acquire language. As Lengeris (2012) states, evidence for the CP hypothesis would require 'a sharp drop-off in a learner's abilities', and 'all early L2 learners should achieve native-like performance' (and vice versa). This is not to say that learners are not still deterred by other aspects like their own L1, but this does highlight the potential that learners could be taught pronunciation, given the right settings.

Aside from the issue of whether or not language learners could ever achieve native-like performance, another question that arises is whether or not there is even a *need* for learners to aim so high. In Munro and Derwing (1999), they discuss the interaction between foreign accent, comprehensibility and intelligibility and point out that the goal for many L2 learners is to communicate and not necessarily sound like a native speaker. They also conduct a study to prove that despite the fact that some speakers may have what some consider a 'heavy accent', that this does not automatically mean that they are unintelligible. They found in their study that errors in prosody tended to affect the speakers' intelligibility the most, which underscores the role of prosody in organizing our utterances.

While linguists make these discoveries and observations of L2 learning, it seems that it takes a lot of effort for them to trickle down to the foreign language classroom. In Darcy et al. (2012), they find through a small survey of 14 teachers that although teachers tend to find pronunciation to be 'very important', the majority do not teach it at all. When asked why they do not teach it, they cited reasons such as 'time, a lack of training and the need for more guidance and institutional support'. Even though the number of teachers surveyed may be significantly small, this gives us a glimpse through the lens of what language teachers themselves experience in relation to pronunciation. We see that even though teachers would like to address it, this would require a restructuring in their curriculum and training– something that would undoubtedly take even more time before students get more pronunciation attention. Compounded with the issue of time and the fact that not all learners need or want equal amount of pronunciation training, it may be unlikely to see such change in second language curriculum so soon.

This points to the potential solution of employing a technology-based system to improve pronunciation as learners could individually address their needs *outside* of the classroom.

3.2 Computer-assisted pronunciation training systems

With the improvements of technology and speech processing, researchers have attempted to make a number of computer-assisted pronunciation training (CAPT) systems. In general, CAPT systems utilize some form of automatic speech recognition (ASR) to record a speaker and compares their recordings (usually) with a native speaker gold standard. They also usually include a feedback mechanism with a combination of pitch contours, spectrograms or audio recordings to help the user adjust their pronunciation.

In Neri et al. (2002), we are presented with an overview of the interaction between language pedagogy and CAPT systems. Here, we see that aside from the classroom, there seems to be an issue in relating the findings of linguistics/language pedagogy with technology. Part of the reason, they suggest, stems from the fact that there are not ‘clear guidelines’ on how to adapt second language acquisition research and thus many CAPT systems ‘fail to meet sound pedagogical requirements’. They emphasize the need for the learners to have appropriate input, output, and feedback and exhibit how the systems available at the time were lacking. For example, they criticize some CAPT systems that were prevalent at the time including systems like *Pro-nunciation* and the *Tell Me More* series for utilizing feedback systems that give the users feedback in waveforms and spectrograms, which cannot be easily interpreted without training. Further, they argue that although visual feedback has its merits, this kind of feedback suggests to the user that their utterance must look close to what is shown on the screen, which is not the case. An utterance can be pronounced perfectly fine, but look completely different from a spectrogram, and *especially* a waveform due to the number of features represented in each visualization, such as the intensity, which will indefinitely vary from user to user and the given examplar. They conclude their article by making it a point to discuss recommendations for CAPT systems, by stating that they should integrate what has been found in research from second language acquisition, and to train pronunciation in a communicative manner to give context to the learners. They also point to the problematic area of feedback and advise that systems provide more easily interpretable feedback with both audio and visual information, and propose that systems give exercises that are ‘realistic, varied, and engaging’. Despite the fact that this article was published in 2002, this article provides a sound basis in addressing the proper makings of a successful CAPT system.

In another article by Eskenazi (2009), we are given a brief review of technologies in CAPT systems, this time more focused from a technical perspective. In particular, the author gives attention to the different CAPT system types and provides information on prosody detection and complete tutoring systems.

The article explains that CAPT systems can be generally split into two main types: individual error detection and pronunciation assessment. As indicated, individual error detection systems are more focused on one particular aspect of the user’s speech, such as the phones or pitch, while pronunciation assessment systems are more designed to represent how a human would judge a non-native utterance.

Early individual error detection systems, including one of her very own Eskenazi and Hansma (1998), started by using a variety of speech recognition techniques such as forced alignment or unconstrained speech recognition. They also worked with a variety of measures to detect the differences between the individual errors and gold standard. Some of these measures include hidden Markov model (HMM) based recognition scoring, a confidence score based system known as Goodness of Pronunciation (GOP), and Linear Discriminant Analysis (LDA). Each of these measures were found to somehow detect the users' errors; however they suffer from issues like low precision or the need for a very homogeneous sample (e.g. Japanese speakers).

Here, Eskenazi (2009) makes a point that working to improve non-native pronunciation is not simply a binary question of native vs. non-native; instead the L1 of the system's users must be considered, as this can greatly affect the evaluation. She also points out that the level of language learning of the speakers can also impact the metrics and success of the system as well, and thus an appropriate population must be selected carefully when building a CAPT system, especially when considering individual errors.

In her discussion of prosody correction, she points to pivotal works that have used a variety of manners to address the issue. Some works include systems that use Pitch Synchronous Overlap and Add (PSOLA) to resynthesize the prosody of users to help them hear what an appropriate utterance would sound like. This in particular could be a potentially effective feedback mechanism to employ in future systems, as it has been said that imitating one's own voice is the most effective. Other systems she mentions include systems that use appropriate L2 phonological models and break prosody down into two levels— syllable-word and utterance-phrase, and systems that detect the 'liveliness' of a speaker. However, she does not discuss prosody correction systems in much detail, which may suggest that there is not as much research in this particular area as compared to the individual error systems. Regardless, these works all provide interesting paths to consider in developing a prosody correction system.

[This part might need to go; replace with accent teaching work?]

Similar to Eskenazi (2009), Chun et al. (2008), presents a review of various technologies, this time related directly to prosody. They discuss four main tools in teaching prosody: 'visualization of pitch contours', 'multimodal tools', 'spectrographic displays' and 'vowel analysis programs'. Citing previous work, it appears that they suggest that the visualization of pitch contours is the most robust method of feedback for learners as it is the most intuitive and non-language specific. Aside from this however, they also discuss the potential of a multimedia approach used by Hardison (2005) that integrate both audio and video in a system called *Anvil*. Following this research, users of this system were able to generalize their training beyond a sentence level and were able to perform better at a discourse-level. This again emphasizes the point that prosody training should put the language in context, which is an important aspect to consider prosody training, as we know how prosody works in relation to communication.

They also discuss the two main methods of such prosody systems: one which utilizes isolated scripted sentences and the other utilizing imitation. They conclude that neither method is useful for generalizing to novel methods and suggest that the training should relate to the ultimate goal. Other information they provide in this article are prosody models used in previous studies. We see that some previous studies have focused on utilizing a variety of sentence types to teach prosody, contrasting *wh*-questions, echo questions, either-or questions and statements. Like the other articles, the works examined in Chun et al. (2008) gives us insight on potential ways to improve future CAPT systems, as we are shown exemplars of potential input and positive reinforcements in successful types of feedback for the user. They conclude that in order to create better pronunciation training systems, we should take advantage of recent technology.

In recent works related to gamified language teaching, (Tejedor-García et al. 2017) experiment with utilizing synthetic voices for corrective feedback in a pronunciation training tool. In their study, they use Google's offline Android text-to-speech (TTS) system as feedback for B1 and B2 Spanish learners of English, and have them focus on the six most difficult pairs of vowels **[insert IPA here?]**. In order to train the users, the researchers first had them watch videos that describe the articulatory/perceptive features of the vowels, and had them listen to a number of minimal pairs produced by the TTS system in succession. Afterwards, they were asked to discriminate minimal pairs in a listening task and then asked to pronounce them.

From this study, they conclude that making use of common commercial TTS are beneficial for users and instructors alike as indicated by both the improvement in performance by the users and the feedback given by those involved in the experiment. This provides further support for works like (Felps et al. 2009), who demonstrated that accent converted speech also However, because the study was limited to individual words and only six pairs of vowels, further experimentation needs to be conducted in order to fully support their claim.

3.3 Voice conversion

Previous works

There have been a number of efforts to design voice conversion systems using various methodologies. One particular method that has been applied recently from other areas of speech technology is the usage of *i-vectors*. I-vectors are akin to word embeddings in text-based natural language processing tasks in the sense that i-vectors encapsulate any type of desired speech information in a vectorized fashion. I-vectors have proven to be successful in a number of tasks, such as speaker verification, language identification, and native accent identification. Of particular interest is DeMarco and Cox (2013), who shows that

In the instance of voice conversion, i-vectors are made of speaker super-vectors trained on GMMs and low dimensional features that represent an individual speaker's features (Wu et al. 2016). This is extracted per utterance and then averaged to form an i-vector that represents an individual speaker. In this way, a source speaker's i-vector can be approximated towards a target speaker's i-vector by a mapping function using neural networks, gaussian mixture models, or other appropriate algorithms.

The usage of i-vectors in voice conversion has been seen in works such as Wu et al. (2016) and Kinnunen et al. (2017). Following Kinnunen et al. (2017), the usage of i-vectors in voice conversion aligns perfectly with the task as it is highly similar to speaker verification; however instead of being a classification task (e.g. is this said speaker or not), voice conversion is a regression task. In Wu et al. (2016), they test and compare a variety of frameworks, such as a deep bi-directional long-short term memory neural (DBLSTM) network architecture, a DBLSTM combined with an average voice model, a DBLSTM combined with an average voice model retrained on some paralleled data, and another model which combined a DBLSTM, average voice model, and i-vectors. In order to evaluate these models, they provide both an objective evaluation using a measure known as mel-cepstral distortion (MCD) and a subjective evaluation rated on quality and similarity, which was decided by the votes of 20 listeners.

[expand on Wu et al. (2016) more]

Following the results of the subjective evaluation, they find that adding i-vectors to the DBLSTM and average voice model outperforms the DBLSTM and average voice model *without* i-vectors and that the DBLSTM and average voice model with i-vectors performs almost as well as the DBLSTM that was retrained. This underscores the need for i-vectors to capture speaker characteristics; without, the system highly underperforms and has poor quality and similarity.

As oppose to Wu et al. (2016) which utilizes an *eigenvoice* (or average voice), Kinnunen et al. (2017) supersedes Wu et al. (2016) by not requiring *any* parallel data.

Although most voice conversion systems have been successful at the general task, many of them suffer from low quality and/or low naturalness in their final outputs. For example, in listening to the audio of Wu et al. (2016)[1] it is apparent that regardless of the low quality of the original source and target audios, the quality of the converted audio sounds muffled. This can be attributed to the various nuanced steps and features required to have high quality voice conversion.

for example, in a shared task dedicated to voice conversion, appropriately called *The Voice Conversion Challenge* where many leading research groups involved in speech technology around the world have submitted systems in attempts to tackle the issue. In the second iteration of the challenge Lorenzo-Trueba et al. (2018), the organizers proposed both a parallel and non-parallel version of the task, both of which were evaluated on natural and similarity using crowdsourcing.

The type of systems submitted to the 2018 edition of the task displays the current state of voice conversion and perhaps machine learning research in general as this year saw a huge increase in the number of systems using neural networks. However, it does not go without saying that there were indeed systems that used more traditional statistical methods, such as Gaussi an Mixture Models (GMM) and one of its variations, differential GMM (DIFFGMM).

In order to evaluate the systems, a group of roughly 300 listeners were gathered to carry out a perceptual evaluation. The systems were evaluated on two main measures: naturalness, which was evaluated on a scale of 1 (completely unnatural) to 5 (completely natural); and similarity, which was evaluated using a same/different paradigm. Following the results, only one system, referred to as N10, was able to outperform the baseline in terms of naturalness (alongside the original source and target audios). When observing the performance of other systems in terms of similarity, we see about 5 our of 23 submitted systems outperforming the baseline. From this, we can conclude that it easier to create a system with high similarity than high naturalness, which is consistent with other common systems.

In discussing the results of the N10 system, the authors credit the success of the system to the *hundreds of hours* of external speech data that was utilized to train a model to recognize content-related features, as well manual fine-tuning. The creators of this system also made use of WaveNet, a novel high-fidelity vocoder and dozens of hours of clean English speech, which could also explain the success of their results. Thus, as previously discussed, we can conclude that creating a high-fidelity voice conversion

[1] Visit http://www.nwpu-aslp.org/vc/apsipa-jiewu-demo.pptx to hear samples.

requires not only appropriate fine-tuning of the data, but also a large amount of external data to support the system.

Thus, even though many systems were neural network based, only one neural network based system was able to outperform the sprocket GMM-based baseline, which could suggest that NN-based methods require proper fine-tuning of the hyperparameters.

Although we see limitations in the systems presented in The 2018 Voice Conversion Challenge, there have bene other efforts to present high quality voice conversion systems in works such as and Nguyen et al. (2016) and Fang et al. (2018).

Fang et al. (2018) leverages a cycle-consistent adversarial network (CycleGAN) architecture, a variation of the recently trending generative adversarial network (GAN) architecture, which was originally used for unpaired image-to-image translation. For example, GANs have been shown to be able to convert images of zebras into horses, as well as winter into summer.

[move this paragraph into accent conversion? discuss this as motivation/inspiration of accent conversion] there have also been some incredible breakthroughs in systems set forth by research teams at Google Brain. One such system involves the Tacotron end-to-end system, which has been proposed to replace the current set-up of text-to-speech systems by reducing the amount of components (decoder, vocoder, etc. [IS THIS TRUE?]) into one piece. The researchers working on this system have recently revealed a impressive system that also takes advantage of deep neural networks to encode speaker characteristics into embeddings, which are then utilized to transfer style (Wang et al. 2018).

With that said, it is evident that the reason for the success of their systems is due to the availability of large-scale, high quality data that many research institutions do not have access to or have funding for. Thus, it may be a long while before the general public has the ability to replicate such systems; however it is extremely exciting to know that there is the possibility.

3.4 Accent conversion

Due to the specialized nature of accent conversion as compared to voice conversion, there are fewer articles and systems available for reference. In fact, most articles that are easily accessible on accent conversion were all published by the same group of researchers at Texas A&M University.

With that said, Aryal and Gutierrez-Osuna (2014b) and other works done by the group of researchers have made efforts to address the challenge. Throughout their research, they test a variety of methodologies, including accent conversion through voice morphing and articulatory synthesis. In the same work of Aryal and Gutierrez-Osuna (2014b), they propose a variation to standard forced alignment techniques used in voice conversion to pair frames based on acoustic similarity. To achieve this, they extract 24 MFCCs per utterance and then dampen the vocal tract differences between the speakers using a method known as vocal tract length normalization (VTLN). They then find the closest frames using clusters, which are then mapped using a GMM conversion method.

In order to evaluate their system, they had a group of 13 participants rate 12 utterances from the test set for their perceived accent (Which utterance was less accented?) and perceived speaker identity (Does utterance X sound more similar to A or B?). This system was compared to a standard voice conversion system that uses standard forced alignment and trained using GMMs. They found that comparing the AC system to the original L2 audio resulted in participants rating the converted audio as sounding less accented 86% of the time, while the VC system compared to the original L2 audio was rated at 91% of the time. However, when the converted audios from both systems were compared, participants rated the AC system to be less accented compared to the VC system 59% of the time. It was also concluded that the AC system was more successful in retaining speaker identity, as the participants found the converted audio more similar to the L2 speaker 78% of the time. More interestingly, they found that the AC system was especially effective in converted utterances that are harder for the L2 speaker to pronounce. This was measured by examining the relationship between the number of phonemes that do not exist in the L2 language (in this case Spanish), and the number of listeners who judged the converted speech as sounding less accented.They found that there was a 0.86 correlation, indicating the robustness of the AC system. Thus, it appears that adjusting the alignment method to align acoustically similar sounds is a good start for accent conversion systems.

[Add other Aryal papers here]

Finally, in Aryal and Gutierrez-Osuna (2015), we see a more novel method that looks beyond acoustic features to perform accent conversion.Citing the results of their previous work, they motivate the usage of articulatory information in accent conversion reasoning that acoustic-based systems often struggle in the challenge of separating accent from speaker identity, which causes the accent converted audio to sound like a combination of the L1 speaker and L2 speaker. To do this, they propose a system that combines both the more standard acoustic information like aperiodicity, pitch and energy from the L1 speaker with articulatory information recording using an electromagnetic articulograph (EMA). Like many recent works, they test a DNN-based mapping function between the L1 and L2 data, which they compare to the previously popular

GMM-based system.

In the evaluation of their system, they use crowdsourced efforts to rate their system based on intelligibility, accentedness, and speaker identity. They find that the DNN-based system was rated to have a 4.3 out of 7 in terms of intelligibility as compared to 3.84 out of 7 for the GMM-based system, proving that DNNs are more robust in this instance. The participants also rated the DNN-based system to be more native-like in 67% of cases as compared to the GMM-based system. With that said, the test set was only 15 sentences, which indicates that 10 out of 15 sentences were better with the DNN system; thus the test set used may be too small to draw hard conclusions. The most important conclusions drawn from their experiments was that of the voice identity assessment. In asking the participants to rate whether an MFCC compression and AC audio from the DNN and GMM-based systems, they found that the participants were fairly confident that the two audios were from the same person with both systems, with the DNN-based system outperforming the GMM-based system as before at a score of 4.3 out of 7 on average, and the GMM-based system at a score of 4.0. However, this is difficult to compare to more common acoustic-only accent conversion systems, as this is not including in their evaluation. With that said, it may be possible to conclude that this would outperform acoustic-based systems, as they proposed this system to tackle flaws in their previous work.

Evidentally, although including articulatory information seems to improve the performance of accent conversion systems, as discussed in the closing remarks of their paper, recording articulatory information can cost a great deal of money and time (Aryal and Gutierrez-Osuna 2015). Most publically (and privatized) speech corpora also do not include articulatory information, meaning that experimenting with it in accent conversion at a broader scale is unfeasibile. Thus, it is ambitious to accept adding articulatory information to accent conversion systems and further work needs to be done in order to scale standard audio-based speech corpora.

Aside from the work done by these researchers, it appears that not much has been done since to address accent conversion. Looking at their recent publications, it seems that they have also halted work in this area, as they have not published articles in the area since 2016; thus this leaves a gap in the research. However, research in voice conversion continues to expand, which leaves the potential of apply methodologies from voice conversion to accent conversion. Following the general methodologies of voice conversion, I hypothesize that it should be plausible to convert accents in a similar fashion and apply more recent innovations to propose state-of-the-art (?!) methods.

Chapter 4

Design and methodology

In this chapter, I introduce the dataset and tools utilized in the experiments, and detail the procedures carried out to conduct the accent conversion process.

4.1 Data

The main dataset utilized in the following experiments is the [put corpus here].

4.2 Experiment 1: GMM-based accent conversion

This experiment a) is to understand more traditional mapping methods used in voice/accent conversion and b) serves as a baseline to be compared to other innovative methods (e.g Experiment 2).

The methodology for this experiment stems from Aryal and Gutierrez-Osuna (2014b), one of the earlier works done on accent conversion.

4.2.1 Tools

In order to do GMM-based accent conversion, I utilize the `nnmnwkii`[1] Python package which provides fast and easy functions to train voice conversion systems. Alongside this package, I also utilize a number of other packages that `nnmnkwii` is dependent on, including `pysptk`, a Python wrapper for the Speech Processing Toolkit, `pyworld`, a Python wrapper for WORLD, a well-known tool for high-quality speech analysis

[1] Found at: https://github.com/r9y9/nnmnkwii

and acoustic feature extraction, `librosa`, another package for audio analysis, and the common `scikit-learn` machine learning package for GMM training.

4.3 Experiment 2: I-vector based accent conversion

This experiment is motivated by the works presented in Wu et al. (2016) and Kinnunen et al. (2017). Due to their flexible nature, i-vectors are an appropriate method to capture the representation of an accent in a compact way.

4.3.1 Tools

In order to do the i-vector based accent conversion, I first utilized the `SIDEKIT` Python toolkit to extract the MFCCs, create a GMM (also known as a universal background model) and then the i-vectors to represent each accent.

Chapter 5

Experimental results

In this chapter, I present the results of the experiments described in the previous chapter and discuss their outcomes and shortcomings.

Bibliography

Aryal, S. & Gutierrez-Osuna, R. (2014a). Accent conversion through cross-speaker articulatory synthesis. (pp. 7694–7698). IEEE. doi:10.1109/ICASSP.2014.6855097

Aryal, S. & Gutierrez-Osuna, R. (2014b). Can voice conversion be used to reduce non-native accents? (pp. 7879–7883). IEEE. doi:10.1109/ICASSP.2014.6855134

Aryal, S. & Gutierrez-Osuna, R. (2015). Articulatory-Based Conversion of Foreign Accents with Deep Neural Networks, 5.

Bongaerts, T., Planken, B., & Schils, E. (1995). Can Late Starters Attain a Native Accent in a Foreign Language? A Test of the Critical Period Hypothesis.

Chun, D. M., Hardison, D. M., & Pennington, M. C. (2008). Technologies for prosody in context: Past and future of L2 research and practice. (pp. 323–346).

Darcy, I., Ewert, D., & Lidster, R. (2012). Bringing pronunciation instruction back into the classroom: An ESL teachers' pronunciation "toolbox", 18.

DeMarco, A. & Cox, S. J. (2013). Native Accent Classification via I-Vectors and Speaker Compensation Fusion, 5.

Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication*, *51*(10), 832–844. doi:10.1016/j.specom.2009.04.005

Eskenazi, M. & Hansma, S. (1998). The Fluency Pronunciation Trainer, 6.

Fang, F., Yamagishi, J., Echizen, I., & Lorenzo-Trueba, J. (2018). High-quality nonparallel voice conversion based on cycle-consistent adversarial network. *arXiv:1804.00425 [cs, eess, stat]*. arXiv: 1804.00425 `[cs, eess, stat]`

Felps, D., Bortfeld, H., & Gutierrez-Osuna, R. (2009). Foreign accent conversion in computer assisted pronunciation training. *Speech Communication*, *51*(10), 920–932. doi:10.1016/j.specom.2008.11.004

Hardison, D. M. (2005). Contextualized Computer-based L2 Prosody Training: Evaluating the Effects of Discourse Context and Video Input. *CALICO Journal*, *22*(2), 16.

Jurafsky, D. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Prentice Hall series in artificial intelligence. Upper Saddle River, N.J.: Pearson Prentice Hall.

Kinnunen, T., Juvela, L., Alku, P., & Yamagishi, J. (2017). Non-parallel voice conversion using i-vector PLDA: Towards unifying speaker verification and transformation. (pp. 5535–5539). IEEE. doi:10.1109/ICASSP.2017.7953215

Lengeris, A. (2012). Prosody and Second Language Teaching: Lessons from L2 Speech Perception and Production Research. In J. Romero-Trillo (Ed.), *Pragmatics and Prosody in English Language Teaching* (Vol. 15, pp. 25–40). Dordrecht: Springer Netherlands. doi:10.1007/978-94-007-3883-6_3

Lenneberg, E. H. (1967). The Biological Foundations of Language. *Hospital Practice*, *2*(12), 59–67. doi:10.1080/21548331.1967.11707799

Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T., & Ling, Z. (2018). The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods. *arXiv:1804.04262 [cs, eess, stat]*. arXiv: 1804.04262 [cs, eess, stat]

Mohammadi, S. H. & Kain, A. (2017). An overview of voice conversion systems. *Speech Communication*, *88*, 65–82. doi:10.1016/j.specom.2017.01.008

Munro, M. J. & Derwing, T. M. (1999). Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. *Language Learning*, *49*, 285–310. doi:10.1111/0023-8333.49.s1.8

Neri, A., Cucchiarini, C., Strik, H., & Boves, L. (2002). The Pedagogy-Technology Interface in Computer Assisted Pronunciation Training. *Computer Assisted Language Learning*, *15*(5), 441–467. doi:10.1076/call.15.5.441.13473

Nguyen, H. Q., Lee, S. W., Tian, X., Dong, M., & Chng, E. S. (2016). High quality voice conversion using prosodic and high-resolution spectral features. *Multimedia Tools and Applications*, *75*(9), 5265–5285. doi:10.1007/s11042-015-3039-x. arXiv: 1512.01809

Scovel, T. (1988). *A time to speak: A psycholinguistic inquiry into the critical period for human speech.*

Sundermann, D., Hoge, H., Bonafonte, A., Ney, H., Black, A., & Narayanan, S. (2006). Text-Independent Voice Conversion Based on Unit Selection. (Vol. 1, pp. I-81-I-84). IEEE. doi:10.1109/ICASSP.2006.1659962

Tejedor-García, C., Escudero, D., González-Ferreras, C., Cámara-Arenas, E., & Cardeñoso-Payo, V. (2017). Evaluating the Efficiency of Synthetic Voice for Providing Corrective Feedback in a Pronunciation Training Tool Based on Minimal Pairs. (pp. 25–29). ISCA. doi:10.21437/SLaTE.2017-5

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R. J., Battenberg, E., Shor, J., ... Saurous, R. A. (2018). Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. *arXiv:1803.09017 [cs, eess]*. arXiv: 1803.09017 [cs, eess]

Wu, J., Wu, Z., & Xie, L. (2016). On the use of I-vectors and average voice model for voice conversion without parallel data. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (pp. 1–6). doi:10.1109/APSIPA.2016.7820901