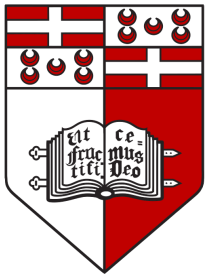
A neural network-based approach to accent conversion

Kenny W. Lino

Msc. 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 De Marco, 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

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.

In this thesis, I aim to investigate 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

Abstract I
Contents III
List of Figures IV
List of Abbreviations IV

1 Introduction **1**
1.1 Research Questions 1
1.2 Thesis Overview 2
1.3 Background and related work 2
1.3.1 Theoretical and educational motivations 2
1.3.2 Computer-assisted pronunciation training systems 3
1.3.3 Voice conversion 6
1.3.4 Accent conversion 7

List of Figures

List of Abbreviations

CAPT	Computer Assisted Pronunciation Training
CP	Critical Period
L1/L2	First and second language

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]. 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 to the same extent. One such example 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. One particular example of something that has been fairly stagnant in language education is pronunciation. Unlike grammar and vocabulary, pronunciation can be challenging to both learn and teach due to the lack of clarity on how to teach it.

1.1 Research Questions

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

- How can we leverage deep neural network technology and voice conversion to convert language learner's speech into sounding more native-like?
- Should we be able to create a sound voice conversion system, would it be possible

to convert the language learner's speech with minimal (non-parallel) audio?

1.2 Thesis Overview

The overview of the thesis is as follows: The main research question of this thesis is the following:

1.3 Background and related work

In this section, I provide a brief overview of second language acquisition and education in order to motivate the usage of technology in language learning using tools such as the one proposed here in this thesis. I then examine some previous research in computer assisted pronunciation (CAPT) systems in order to frame the successes and gaps of such work, and close with a discussion about voice conversion and accent conversion.

1.3.1 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.

1.3.2 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, she gives attention to the different CAPT system types and provides information on prosody detection and complete tutoring systems.

She first 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.

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.

1.3.3 Voice conversion

Similar to CAPT systems, many researchers have steadily progressed on building voice conversion systems. However, unlike CAPT systems, VC systems are not typically designed for uses in language learning and are more grounded in other speech technology uses such as text-to-speech synthesis.

To properly frame voice conversion, we take a look at Mohammadi and Kain 2017 who presents 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 *target speaker* to make it sound similar to a *source speaker* in any chosen fashion with the utterance still being intact Mohammadi and Kain 2017. 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 information is different, while non-parallel data would indicate datasets where the utterances are not the same.

Even though work in VC has progressed within the last decade or so, it seems that the task still presents a large challenge for researchers due to the various nuanced steps and features required to have high quality voice conversion. This can be witnessed

for example, in a shared task dedicated to voice conversion, appropriately called *The Voice Conversion Challenge* where many 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 this year's version of the task displays the current state of the field 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 Gaussian Mixture Models (GMM) and one of its variations, differential GMM (DIFFGMM).

In order to evaluate the systems, a group of listeners roughly 300 were gathered using crowdsourcing. We

Thus, even though not many systems were neural network based, only one neural network based system was able to outperform the sprocket GMM-based baseline.

Although we see limitations in the systems presented in The 2018 Voice Conversion Challenge, 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.

1.3.4 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).

Finding previous work done in accent conversion has proven to be a difficult task as there are not many articles available in the area; this may be because work in voice conversion itself is already a subfield of speech technology.

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 do achieve this, they first dampen vocal tract differences between the speakers using a method known as *vocal tract length normalization*.

Finally, in Aryal and Gutierrez-Osuna 2015, they also join in on the rising trend of utilizing deep neural networks. However, instead of training on solely audio, they utilize articulatory information for real-time conversion.

Evidentally, an articulatory-based accent conversion system requires specialized technology that is inaccessible to most users, and thus such a system cannot be adopted easily.

Aside from the work done by these researchers, 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 between the potential for accent conversion and language learners.

Bibliography

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

Aryal, S. & Gutierrez-Osuna, R. (2014b, May). 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.

Eskenazi, M. (2009, October). 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.

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

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, December 1). 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, April 11). The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods. arXiv: 1804.04262 [cs, eess, stat]. Retrieved April 13, 2018, from http://arxiv.org/abs/1804.04262

Mohammadi, S. H. & Kain, A. (2017, April). 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, December 1). 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

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

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R. J., Battenberg, E., Shor, J., … Saurous, R. A. (2018, March 23). Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. arXiv: 1803.09017 [cs, eess]. Retrieved April 3, 2018, from http://arxiv.org/abs/1803.09017