STATISTICAL PARAMETRIC METHODS FOR ARTICULATORY-BASED FOREIGN ACCENT CONVERSION

A Dissertation

by

SANDESH ARYAL

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Ricardo Gutierrez-Osuna
Committee Members,	Yoonsuck Choe
	Dylan Shell
	Byung-Jun Yoon
Head of Department,	Dilma Da Silva

December 2015

Major Subject: Computer Engineering

ABSTRACT

Foreign accent conversion seeks to transform utterances from a non-native speaker (L2) to appear as if they had been produced by the same speaker but with a native (L1) accent. Such accent-modified utterances have been suggested to be effective in pronunciation training for adult second language learners. Accent modification involves separating the linguistic gestures and voice-quality cues from the L1 and L2 utterances, then transposing them across the two speakers. However, because of the complex interaction between these two sources of information, their separation in the acoustic domain is not straightforward. As a result, vocoding approaches to accent conversion results in a voice that is different from both the L1 and L2 speakers. In contrast, separation in the articulatory domain is straightforward since linguistic gestures are readily available via articulatory data. However, because of the difficulty in collecting articulatory data, conventional synthesis techniques based on unit selection are ill-suited for accent conversion given the small size of articulatory corpora and the inability to interpolate missing native sounds in L2 corpus.

To address these issues, this dissertation presents two statistical parametric methods to accent conversion that operate in the acoustic and articulatory domains, respectively. The acoustic method uses a cross-speaker statistical mapping to generate L2 acoustic features from the trajectories of L1 acoustic features in a reference utterance. Our results show significant reductions in the perceived non-native accents compared to the corresponding L2 utterance. The results also show a strong voice-similarity between accent conversions and the original L2 utterance. Our second (articulatory-based) approach consists of building a statistical parametric articulatory synthesizer for a non-native speaker, then driving the synthesizer with the articulators from the reference L1 speaker. This statistical approach not only has low data requirements but also has the flexibility to interpolate missing sounds in the L2 corpus. In a series of listening tests, articulatory accent conversions were rated more intelligible and less accented than their L2 counterparts. In the final study, we compare the two approaches: acoustic and articulatory. Our results show that the articulatory approach, despite the direct access to the native linguistic gestures, is less effective in reducing perceived non-native accents than the acoustic approach.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Dr. Ricardo Gutierrez-Osuna for his continuous guidance, encouragement and support throughout my graduate studies that helped me become a better researcher. I would also like to thank my committee members Dr. Yoonsuck Choe, Dr. Dylan Shell, and Dr. Byung-Jun Yoon for the valuable suggestions and feedbacks. I am also grateful to all my teachers who contributed in making me ready for undertaking this path.

Thanks also go to my lab mates Jin, Avinash, Daniel, Jong, Rakesh, Folami, Rhushabh, Virendra, Chris, Anshul, Zelun, Difan, Adam and Guanlong for keeping my spirits up, being always ready to help whenever needed, and making my time in the lab enjoyable.

Finally, I am very grateful to my parents and my family for their inspiration, encouragement, patience, unwavering support and love. A special gratitude goes to my wife Sushila for being there when I needed the most.

TABLE OF CONTENTS

Page

ABSTRACT ii

ACKNOWLEDGEMENTS iv

TABLE OF CONTENTS v

LIST OF FIGURES ix

LIST OF TABLES xiii

1. INTRODUCTION 1

1.1 Thesis outline 6

2. BACKGROUND 7

2.1 Non-native accents 8
2.1.1 Non-native accents in communication 10
2.1.2 Pronunciation training for second language learner 11
2.2 Speech production physiology 13
2.2.1 Acoustic theory of speech 14
2.2.2 Speaker's identity, voice and accent 16
2.2.3 Modulation theory 17
2.3 Speech analysis and synthesis 18
2.3.1 Linear predictive analysis 18
2.3.2 Cepstral analysis 20
2.3.3 STRAIGHT analysis 21
2.3.4 Sinusoidal analysis and harmonics plus noise model 22
2.4 Articulatory speech processing 23
2.4.1 Articulatory representation 24
2.4.2 Articulatory measurements 25
2.4.3 Articulatory normalization 27
2.4.4 Articulatory synthesis 30
2.4.5 Articulatory inversion 33
2.5 Summary 35

Page

3. LITERATURE REVIEW 36

3.1 Foreign accent conversion 37
3.2 Acoustic-based approach 37
3.3 Articulatory-based approaches 42
3.4 Evaluation of foreign accent conversion 45
3.4.1 Acoustic quality assessment 46
3.4.2 Intelligibility assessment 47
3.4.3 Assessment of non-native accents 47
3.4.4 Assessment of voice-identity 49

4. ACOUSTIC-BASED FOREIGN ACCENT CONVERSION USING VOICE CONVERSION 50

4.1 Foreign accent conversion based on spectral mapping 51
4.1.1 STRAIGHT feature extraction and synthesis 53
4.1.2 Pairing acoustic vectors 53
4.1.3 Cross-speaker spectral mapping 55
4.1.4 Prosody modification 59
4.2 Experimental 60
4.2.1 Conversion from non-native to native accent 60
4.2.2 Conversion from native to non-native accent 61
4.2.3 Experimental corpus 62
4.3 Results 63
4.3.1 L1→L2 accent/voice conversion 63
4.3.2 L2→L1 accent/voice conversion 64
4.3.3 Correlation with differences in the L1 and L2 phonetic inventories 65
4.4 Discussion 67

5. STATISTICAL PARAMETRIC ARTICULATORY FOREIGN ACCENT CONVERSION 69

5.1 Method description 71
5.1.1 Cross-speaker articulatory mapping 71
5.1.2 Forward mapping 73
5.1.3 System diagram 75
5.2 Experimental corpus 76
5.2.1 Experimental conditions 77
5.2.2 Participant recruitment 80
5.3 Results 81
5.3.1 Accuracy of articulatory normalization 81

Page

5.3.2 Assessment of intelligibility 84
5.3.3 Assessment of non-native accentedness 90
5.3.4 Assessment of voice individuality 92
5.4 Discussion 94

6. ARTICULATORY-BASED CONVERSION OF FOREIGN ACCENTS WITH DEEP NEURAL NETWORKS 97

6.1 Deep neural network in articulatory-acoustic mappings 98
6.2 DNN-based foreign accent conversion method 99
6.2.1 DNN-based forward mapping 100
6.2.2 Global variance adjustment 101
6.3 Performance of DNN-based forward mapping 102
6.3.1 GMM-based baseline methods 102
6.3.2 Experimental 104
6.3.3 Experiment 1: Comparison of DNN vs. GMM 108
6.3.4 Experiment 2: Context length 109
6.3.5 Experiment 3: Network depth 110
6.3.6 Experiment 4: Synthesis time 111
6.3.7 Experiment 5: Subjective assessment 112
6.3.8 Discussions on the performance of DNN-based forward mapping 113
6.4 Evaluation of foreign accent conversion with DNN 115
6.5 Results 116
6.5.1 Intelligibility assessment 116
6.5.2 Assessment of non-native accentedness 117
6.5.3 Voice identity assessment 119
6.6 Conclusion 121

7. ACOUSTIC VS. ARTICULATORY-BASED STRATEGIES 122

7.1 Comparison between the articulatory and acoustic-based strategies 123
7.2 Equivalent articulatory synthesizer for the acoustic-based strategy 124
7.2.1 Training the cross-speaker forward mapping 125
7.3 Experimental validation 126
7.3.1 Experimental conditions 127
7.4 Results 127
7.4.1 Non-native accent evaluation 128
7.4.2 Intelligibility assessment 129
7.4.3 Voice identity assessment 130
7.5 Discussions 132

Page

8. CONCLUSIONS 135

8.1 Summary 135
8.2 Main contributions 137
8.3 Future Work 138
8.3.1 Large scale validation 138
8.3.2 Performance improvement 138
8.3.3 Application of foreign accent conversion methods in computer aided pronunciation training 140
8.3.4 Extension to other articulatory speech modification problems 141

REFERENCES 142

APPENDIX A: FORWARD MAPPING WITH DEEP NETWORKS 154

APPENDIX B: MECHANICAL TURK TESTS SAMPLES 161

APPENDIX C: LIST OF PUBLICATIONS 167

APPENDIX D: PSI STATFAC TOOLBOX 169

LIST OF FIGURES

Page

Figure 1: Human speech production physiology. 14

Figure 2: Acoustic theory of speech: speech signal (right) is the convolution of the glottal source excitation signal (left) and the vocal tract filter response (middle). 15

Figure 3: A simplified computational model of speech production physiology. 16

Figure 4: A typical FFT and LPC spectrum of a nasal speech segment. 19

Figure 5: STRAIGHT analysis and synthesis. 22

Figure 6: Position of the 6 EMA pellets used in our study; UL: upper lip; LL: lower lip; LI: lower incisor; TT: tongue tip; TB: tongue blade; TD: tongue dorsum. An additional pellet (red cross-hair) was placed on the upper incisor and served as a reference. 27

Figure 7: Cepstral decomposition of speech into spectral slope and spectral detail (DCT: Discrete cosine transform). 41

Figure 8: Articulatory foreign accent conversion based on unit selection (from Felps (2011). 44

Figure 9: Foreign accent conversion method using cross-speaker statistical mappings. 52

Figure 10: (a) Conventional approach to voice conversion; source and target utterances are paired based on their ordering in a forced-aligned parallel corpus, (b) Our approach to accent conversion: source and target utterances are paired based on their acoustic similarity following vocal-tract-length normalization (VTLN), MCD: Mel Cepstral Distortion. 54

Figure 11: Shifting and scaling L1 pitch trajectory to match the vocal range of the L2 speaker. 60

Figure 12: The number of missing phonemes in L2 inventory ($Np \notin L2$) and the proportion of listeners who found the AC12 synthesis less foreign accented than the VC12 synthesis for each test sentence are highly correlated ($\rho = 0.74$). 66

Figure 13: Articulatory accent conversion is a two-step process consisting of L1-L2 articulatory normalization and L2 forward mapping. 70

Figure 14: Overview of the cross-speaker articulatory normalization procedure. A separate set of parameters is obtained for each EMA pellet. 72

Figure 15: Block diagram of accent conversion method (PM: pitch modification). 76

Figure 16: (a) Distribution of six EMA pellet positions from the L1 speaker (solid markers) and L2 speaker (hollow markers) from a parallel corpus. Large differences can be seen in the span of the measured positions of articulators (UL: upper lip; LL: lower lip; LI: lower incisor; TT: tongue tip; TB: tongue blade; and TD: tongue dorsum). The upper incisor (UI) was used as a reference point. (b) Distribution of EMA pellet positions for the L1 speaker (solid markers) and L2 speaker (hollow markers) following articulatory normalization. 82

Figure 17: Trajectory of the tongue-tip pellet in L1 and L2 utterances of the word ‘that’. The L1 trajectory normalized to the L2 articulatory space is also shown. Arrows indicate the direction of trajectories. 83

Figure 18: (a) Distribution of tongue tip position in frontal vowels for the L1 speaker (dark ellipses) and L2 speaker (light) speaker; ellipses represent the half-sigma contour of the distribution for each vowel. (b) Distribution of tongue tip position in frontal vowels for the L1 speaker after articulatory mapping (dark) and the L2 speaker (light). 84

Figure 19: Box plot of (a) word accuracy and (b) subjective intelligibility ratings for $L1_{EMA}$, $L2_{EMA}$and AC utterances. 85

Figure 20: Word accuracy for AC and $L2_{EMA}$ for the 46 test sentences. The diagonal dashed line represents The sentences for which $W_{acc}(AC) > W_{acc}(L2_{EMA})$ are above the dashed line and the vice versa. 86

Figure 21: Subjective evaluation of non-native accentedness. Participants were asked to determine which utterance in a pair was more native-like. 91

Figure 22: Average pairwise voice similarity scores. Scores range from -7 (different speaker with high confidence) to +7 (same speaker with high confidence). 93

Figure 23: (a) DNN-based foreign accent conversion (PM: pitch modification) (b) Forward mapping using a DNN with a tapped-delay line input. 100

Figure 24. Signal processing flow during articulatory synthesis. 105

Figure 25: Trajectories of selected acoustic and articulatory features from a typical test utterance. The top plot shows the second MFCC predicted by the DNN, sGMM and dGMM alongside the target trajectory extracted from the audio recording of the same sentence. The bottom plots show the trajectories of a few articulatory input features for the same utterance. TT_x: anteroposterior position of the tongue tip, TT_y: height of the tongue tip, LL_y: height of the upper lip. 107

Figure 26: (a) Experiment 1: Mel cepstral distortion (MCD) for the DNN, sGMM and dGMM mappings. (b) Experiment 2: MCD for the DNN and GMM as a function of the input articulatory context window. (c) Experiment 3: MCD for the DNN as a function of the number of hidden layers; error bars denote standard errors of means. 109

Figure 27: (a) Experiment 4: Synthesis time of a DNN mapping increases with the size of the network. (b) Experiment 5: Pairwise comparison between DNN and dGMM synthesis; error bars denote standard errors of means...112

Figure 28: (a) Word accuracy and (b) subjective intelligibility ratings for AC_{DNN} and AC_{GMM}. 117

Figure 29: Subjective evaluation of accentedness. Participants selected the most native-like utterances (a) between AC_{DNN} vs. L2 articulatory resynthesis, and (b) between AC_{DNN} vs. AC_{GMM}. 118

Figure 30: Average pairwise voice similarity scores. $L2_{MFCC}$& AC_{GMM} * from section 5.3.4. 120

Figure 31: (a) cross-speaker spectral mapping for acoustic-based accent conversion, and (b) a cross-speaker forward mapping for the equivalent articulatory synthesis of acoustic-based accent conversion. 125

Figure 32: Two-step process for building a cross-speaker forward mapping, $f\colon \boldsymbol{x}_{L1} \rightarrow \boldsymbol{y}_{L2}$ 126

Figure 33: Subjective evaluation of accentedness. Participants selected the most native-like utterances between AC_{EQV} vs. AC_{EMA} 128

Figure 34: (a) Word accuracy and (b) subjective intelligibility ratings for AC_{EQV} and AC_{EMA} 130

Figure 35: Average pairwise voice similarity scores. $L2_{MFCC}$ & $AC_{EMA}{}^{*}$ from section 6.5.3. 132

Figure 36: (a) Forward mapping via a deep neural network (DNN) with a tapped-delay line input. (b) The Gaussian-Bernoulli deep Boltzmann machine as an undirected graphical model with real valued visible units $\boldsymbol{v}$ and binary hidden units $\boldsymbol{h}$. 155

Figure 37: The pre-qualification test to identify the American accents. This qualification test was used to select the native speakers of American English for the perceptual listening tests. 162

Figure 38: The intelligibility assessment test. The participants were asked to transcribe the audio clips and rate their intelligibility. 163

Figure 39: Forced pairwise comparison of non-native accentedness. The participants listened to the two clips of the same sentence and selected the one that sounded the most native-like. Participants were also asked to transcribe the sentence to ensure that they listen to the sentence. 164

Figure 40: The voice-similarity test asks the participants to listen to a pair of linguistically different utterances separated by a beep and answer (i) if the utterances were from the same speaker or not, and (ii) how confident they are on their decision. 165

Figure 41: Subjective evaluation of the acoustic quality. The participants were asked to rate the utterances in the MOS scale (Bad:1, Poor:2, Fair:3 Good:4, Excellent:5). Rated samples were provided for the reference. 166

LIST OF TABLES

Page

Table 1: Sixteen articulatory parameters necessary and sufficient to characterize all possible sounds in human languages (Ladefoged, 1980). 26

Table 2: Summary of the six synthesis models (AP: aperiodicity from STRAIGHT; DTW: dynamic time warping; AC: accent conversion; VC: voice conversion; '12' denotes transformation from L1 to L2). 62

Table 3: Summary of perceptual results; percentage denotes preference for the first item in the pair (second item in ABX); SA: source accent; ID: speaker identity. 65

Table 4: Five experimental conditions for listening test. 78

Table 5: Correlation between word accuracy and the proportion of phones in a sentence containing a particular articulatory-phonetic feature. 89

Table 6: Experimental conditions for the listening tests. 116

1. INTRODUCTION

A non-native speaker who learns to speak a second language after a “critical age” (Lenneberg, 1967) usually speaks with a foreign accent —a systematic deviation from the phonetic and prosodic norms of the native speech. In many cases, foreign accents lower the intelligibility of the speech (Munro and Derwing, 1995), but even when the intelligibility is not compromised, foreign-accented speakers may be subjected to discriminatory attitude (Giles, 1970; Kalin and Rayko, 1978; Rubin and Smith, 1990). Thus, by improving their pronunciation, adult second language learners have more to gain than mere intelligibility. A common pronunciation training approach consists of repeating after a native speaker. However, several studies have suggested that choosing a suitable target voice to imitate, a so called “golden speaker,” can be more effective for pronunciation training (Nagano and Ozawa, 1990; Probst *et al.*, 2002; Bissiri *et al.*, 2006), the rationale being that by removing all voice-specific differences the learner can focus only on the accent related differences. Because finding such a native speaker for each learner is not practical, Felps *et al.* (2009) suggested using speech modification methods to provide the ideal “golden speaker” for each learner: their own voice, but with a native accent. This dissertation focuses on developing such speech modification methods, which we will refer to as **foreign accent conversion**.

Foreign accent conversion can be performed both in the acoustic and in the articulatory domain. In the acoustic-domain, existing vocoding approaches seek to separate the linguistic (accents) and voice-identity information from a pair of time-aligned utterances of a native and a non-native speaker, then transpose them across the speakers (Felps *et al.*, 2009; Aryal and Gutierrez-Osuna, 2013). However, these two sources of information are convolved in a complex way, and therefore are difficult to decouple when analyzing the acoustic signal. As a result, vocoding often results in accent conversions with the voice of a 'third speaker,' one that is different from either speaker (Felps *et al.*, 2009). Unlike the vocoding-based acoustic methods, foreign accent conversion in the articulatory domain is inherently immune to the 'third-speaker' problem due to the voice-independent representation of linguistic gestures via articulatory data (Traunmüller, 1994). In the only existing articulatory-based method prior to this dissertation, Felps *et al.* (2012) showed that accent conversion can be performed by driving an articulatory synthesizer for the non-native speaker based on unit-selection (Hunt and Black, 1996) using articulatory gestures from a reference native utterance. However, the study reported only a moderate reduction in non-native accents and inconsistent acoustic quality due to the small size of the articulatory-acoustic corpus and the inability of unit selection to produce sounds that do not already exists in the corpus.

Given the limitations of the existing methods, this dissertation investigates strategies in both acoustic and articulatory domains. In the acoustic domain, we propose a statistical mapping approach to estimate equivalent trajectories of acoustic parameters for the non-native (L2) speaker from a reference native (L1) utterance, while avoiding the difficulty of separating and transposing the sources of voice-identity and the linguistic information across the speakers. Statistical mappings of acoustic features from a source to a target speaker have been effectively used in voice conversion, a closely related problem where the objective is to modify speech from a source speaker to match the voice of a target speaker (Kain and Macon, 1998; Stylianou *et al.*, 1998; Toda *et al.*, 2007; Desai *et al.*, 2010). However, the mappings in conventional voice conversion are trained on a set of acoustic feature vectors from the source and the target speaker paired based on their ordering within a parallel corpus. Thus, the mappings are likely to learn the accent-related differences too. We hypothesize that the mapping of accent-related characteristics can be avoided by modifying the conventional training process of voice conversion by pairing the frames from the L1 speakers with that of L2 speaker based on their linguistic-similarity.

Similarly, in the articulatory domain, we propose using statistical parametric articulatory synthesizers (Toda *et al.*, 2008). Unlike unit-selection, statistical parametric synthesis has low data requirement and the flexibility to interpolate new sounds that do not exist in the L2 corpus.

The main objectives of the dissertation are:

i) to investigate the effectiveness of the proposed acoustic-based method (cross-speaker statistical mapping) in reducing the perceived non-native accents,

ii) to investigate the effectiveness of the proposed articulatory-based method (statistical parametric articulatory synthesis) in reducing the non-native accents, and

iii) to compare the performance of acoustic-based and articulatory-based strategies for accent conversion.

This work has three major contributions. First, it presents an acoustic-based foreign accent conversion method free from (i) ‘third speaker’ issues, and (ii) the challenging problems of force-aligning L1 and L2 utterances. Second, it proposes a new articulatory-based method for foreign accent conversion that consists of driving a parametric articulatory synthesizer for the L2 speaker articulators from a reference L1 speaker. Unlike the prior articulatory method (based on unit-selection), the proposed method has low data requirements and the flexibility to interpolate new sounds which a L2 may not produce. More specifically, we explore two articulatory synthesis models based on (i) Gaussian mixture models (GMM), and (ii) deep neural networks (DNN). The GMM-based synthesizer is explored because of its proven performance (Toda *et al.*, 2008). However, it is unsuited for the real-time conversion because of a computationally expensive trajectory-optimization required to reduce spectral discontinuities. Therefore, we also propose a DNN-based synthesizer that avoids the need for such costly trajectory optimization (and reduce the run-time computation costs) by exploiting the temporal nature of speech via contextualized input. This study also evaluates the performance of the DNN-based synthesizer in foreign accent conversion. The third and final contribution is an experimental comparison between the two strategies: acoustic-based and articulatory-based. Since the articulatory-based strategy uses articulatory synthesizer which results in lower acoustic quality synthesis compared to the acoustic-based method, the direct comparison of non-native accentedness between the two accent conversions is biased (Felps and Gutierrez-Osuna, 2010). To account for the quality bias, we develop a method to generate the equivalent articulatory synthesis of the acoustic-based accent conversion strategy.

The contributions made in this dissertation work have been published in several conferences and journals over the past three years. Specifically, the acoustic-based method using cross-speaker spectral mappings was presented at IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) in 2014(Aryal and Gutierrez-Osuna, 2014b). A description of the GMM-based articulatory approach and its performance in foreign accent conversion is published in The Journal of Acoustical Society of America in 2015 (Aryal and Gutierrez-Osuna, 2015b). The DNN-based articulatory synthesizer was published at Computer Speech and Language (Aryal and Gutierrez-Osuna, 2015 (in press)). The performance of the proposed DNN-based articulatory synthesizer in foreign accent conversion was presented at Annual Conference of International Speech Communication Association (INTERSPEECH) in 2015 (Aryal and Gutierrez-Osuna, 2015a). Additional preliminary works not included in this dissertation were also published in different conferences. For example, a vocoding approach for foreign accent conversion was published at INTERSPEECH in 2013 (Aryal *et al.*, 2013). The method uses the spectral slope and the details as the voice identity and linguistic components respectively. Similarly, a preliminary work on the articulatory synthesis driven by the articulators inverted from acoustic features is presented at ICASSP in 2013 (Aryal and Gutierrez-Osuna, 2013). Another work worth mentioning here is an articulatory-based accent conversion method published at ICASSP in 2014, which avoids the need to record articulatory data from the L2 speaker by implementing a cross-speaker forward mapping (Aryal and Gutierrez-Osuna, 2014a).

1.1 Thesis outline

The remaining sections in this dissertation are organized as follows. Section 2 provides background information on foreign accents and relevant speech processing methods for accent conversion. Section 3 reviews the existing foreign accent conversion methods and their limitations. Section 4 describes the statistical mapping technique for foreign accent conversion in acoustic domain and evaluates its effect on reduction of non-native accents compared to conventional voice conversion. Section 5 explains the articulatory-based method for foreign accent conversion using Gaussian mixture model and its effectiveness in reducing non-native accents. In section 6, we present the DNN-based approach for the real-time conversion. Section 7 compares the acoustic-based and articulatory-based strategies we developed for foreign accent conversion. Finally, section 8 concludes with a summary of this dissertation work and its future extensions.

2. BACKGROUND[1]

Speech is a longitudinal compression wave, whose main purpose is to transfer linguistic information. In addition, speech also contains information such as identity, emotional state, accent, gender and age of the speaker. Speech processing methods seek to model the speech signal into parametric representations that capture these characteristics for identification and modification purposes. Among these several characteristics, this dissertation focuses on modifying accents, specifically, modifying non-native utterances to sound more native while preserving the identity of the non-native speaker, a problem that is known as foreign accent conversion.

In this section, we review several topics relevant to the problem of foreign accent conversion. First, we discuss non-native accents and how they affect communication. Second, we review speech production physiology in human and the relevant theories that explain the physiological origin of speech characteristics such as accent and voice identity. Third, we review speech analysis and synthesis techniques, which provide a platform to modify speech characteristics. Finally, we also review relevant topics in articulatory speech processing as our focus is on developing articulatory methods for foreign accent conversion.

[1] The subsections on articulatory speech processing are reprinted with permission from "Reduction of non-native accents through statistical parametric articulatory synthesis," by Aryal and Gutierrez-Osuna, 2015. *J. Acoust. Soc. Am.*, 137, pp. 433-446. ©2015 Acoustical Society of America.

2.1 Non-native accents

Speakers who start to speak a second language (L2) after a certain age—the so-called "critical period" (Lenneberg, 1967; Scovel, 1969)—rarely acquire native-like pronunciation. The systematic deviation from the expected norms of a spoken language observed in such learners is known as foreign accent. The deviations can be observed in the choice of vocabulary, intonation, stress, or as the substitution, deletion or insertion of phones. Highlighting these characteristics in non-native speech, Jenner (1976) defines foreign accent as the "complex of interlingual or idiosyncratic phonological, prosodic and paralinguistic systems which characterizes a speaker of a foreign language as non-native."

The cause for foreign accents has been a long-standing research question among linguists. Lenneberg (1967) and Scovel (1969) suggest that beyond the so-called critical period, which is suspected to run between two to the age of puberty, the brain loses its plasticity and the functional differentiation on the brain reaches completion. The inability to distinguish novel phones in adult second language learners makes it harder for them to produce the correct phones. Along the same line of reasoning, Kempen (1992) suggests that adults lose the knack of listening to speech sounds in isolation, focusing more on the higher semantic and textual level. The lower level phonological activities are, for the better part, automatized. The speech perception after a certain age is in the semantic level rather than the low level of phonetic sounds. Hence, the adult second language learners have difficulty in detecting and producing a new speech sound after a certain age. As a result, foreign accents are highly influenced by the differences in phonetic inventory between the learner’s primary and the second languages (Goldstein, 2001; Helman, 2004; You *et al.*, 2005).

As an example, we discuss how the phonological differences between Spanish and English manifest into the common characteristics of Spanish accented English. Several English consonants and vowels such as /z/, /ʃ/, /ʒ/, /dʒ/, /ɑ/, /ʌ/, and /æ/ do not exist in Spanish. Therefore, Spanish speakers tend to map these phones to the closest Spanish consonants. Similarly, English phones /b/ and /v/ are allophones in Spanish, thus, Spanish speakers often incorrectly substitute them. Furthermore, Spanish phonology allows only a few consonants (e.g. /θ/, /s/, /n/, /r/ and /l/) at the end of the word; therefore, many consonants at the end of the word are often dropped or mispronounced. For example, phone /ŋ/ is usually replaced by /n/ at the end of the word.

The phonology of a language also dictates what kinds of consonants clusters (phonotactics) are allowed. As an effect, Spanish speakers tend to oversimplify some consonants clusters in English that does not exist in Spanish phonotactics (e.g., dropping /t/ at the end of the word ‘twist’). These languages also have significant prosodic differences. For example, in Spanish, the nuclear tone falls on the last stressed syllable in the sentence, which is not always the case in English. In addition, Spanish is syllable-timed (syllables are of equal duration) whereas English is stress-timed (durations between stressed consecutive syllables are equal). Thus, the intonation and rhythm in Spanish-accented English differs significantly from that of native English.

The influence of a speaker’s native language in his production of the second language is so strong that the speaker’s mother tongue can be identified with a high accuracy. For example, Hansen and Arslan (1995) trained word HMMs for different accents and used them to identify accents. Focusing on the specific words that are sensitive to accents, the authors were able to recognize one of the four foreign accents (English, Turkish, German, and Chinese) with more than 93% accuracy.

2.1.1 Non-native accents in communication

Adult learners of a second language sometimes have difficulty making themselves understood. These learners often speak with distinct non-native accents, but the low-intelligibility is not necessarily due to their accent. In a study, Munro and Derwing (1995) reported cases where non-native utterances that had been rated as heavily accented were nonetheless transcribed perfectly by native speakers. While many (Abercrombie, 1949; Crawford, 1987; Morley, 1991) argue that a comfortably intelligible pronunciation is sufficient for second language learners, the negative attitude towards speakers with non-native accents cannot be ignored.

The speakers with non-native accents are frequently received with indifference and subjected to disrespect or discriminatory attitudes towards them. Studies have shown that speakers with non-native accents are perceived as less intelligent (Campbell-Kibler, 2009) and less trustworthy (Ryan and Carranza, 1975; Brennan and Brennan, 1981). They are also susceptible to negative stereotyping related to their perceived ethnicity and socio-economic classes (Gluszek and Dovidio, 2010; Dovidio and Fiske, 2012). The intolerance for foreign accents among employers (Kalin and Rayko, 1978; Sato, 1991) is even more destructive as it directly impacts the livelihood of a person. Thus, non-native learners have more to gain than just the intelligibility by acquiring a more native-like pronunciation.

2.1.2 Pronunciation training for second language learner

Adult second language learners have difficulty losing the non-native accent despite their immersion in the language for a long time. However, studies show that it can be reduced significantly through pronunciation trainings. In a study by Neufeld (1978), a number of learners were subjected to an 18-hour course of pronunciation training. Three native speakers judged the learners' imitations of sample sentences after training, and found half of them as native or near native. In another study, Abrahamsson and Hyltenstam (2008) investigated the proficiency and language aptitude of 42 near-native L2 speakers of Swedish, and found that adult learners with a high degree of language learning aptitude can reach the proficiency of a native speaker.

Regardless of the learner's motivation level, pronunciation training techniques also significantly affect the learning process. The "listen-and-repeat" approach is commonly used in pronunciation training (Nagano and Ozawa, 1990; Probst *et al.*, 2002; Bissiri *et al.*, 2006). However, studies have suggested that the similarity in voice between the teacher and the learner effectively improves learning, the rationale being that by removing all other differences between the reference target utterance and the learner's own production, the learner may focus only on accents-related differences. For example, Nagano and Ozawa (1990) compared two types of training utterances for teaching English pronunciation to Japanese learners. One group mimicked utterances from a reference English speaker whereas the other group repeated after their own previous recording modified to match the prosody of the reference English speaker. The authors found that the group that trained on their own modified utterances improved more than the other group. Similar effects were also observed in prosody training for Italian speakers learning to speak German (Bissiri *et al.*, 2006). Such effects of similarity in voice between the teacher and the learner are not limited to prosody training; similar effects had been observed also in training of segmental characteristics such as vowel quality. For example, (Repp and Williams, 1987) found that the speakers were more accurate when imitating isolated vowel in /u/-/i/ and /i/-/æ/ continua when imitating their own previous production than when imitating those produced by a speech synthesizer.

Instead of modifying the learner's own utterance, Probst *et al.* (2002) investigated the effect of learner-teacher voice similarity in pronunciation training by pairing each learner with a teacher with similar voice quality. In their study, learners were divided into three groups based on the teacher's voice: In these three groups, the teachers were (i) selected randomly, (ii) assigned based on their voice similarity to the learner, and (iii) selected by the learner, respectively. The study found that the group that repeated after the teacher with voice similar to their own —termed as the "golden-speaker"— was able to produce the most native like utterances after the training. Finding such "golden-speaker" for each learner is not practical. Thus, Felps *et al.* (2009) suggested using speech modification technique to generate the ideal "golden-speaker", i.e. the learner's own speech but with a native accent. Such speech modification, also known as foreign accent conversion, is the topic of this dissertation work.

2.2 Speech production physiology

Speech production involves several organs (shown in Figure 1) including lungs, vocal cords, oral cavity, nasal cavity, velum, lips, and jaws, etc. A coordinated effort of these organs produces variations in the air pressure which is perceived as sound. In the case of voiced sounds, such as /b/ and /d/, the lungs pump the air out through trachea causing the vocal cords at the glottis to vibrate. The pulsating airwave then passes through the vocal tract and comes out of the mouth; the resonance in the vocal tract modulates the vibration giving specific linguistic characteristics to the sound. The articulators such as tongue, lips, velum and jaws are responsible for generating different sounds by altering the vocal tract configuration. In the case of unvoiced sounds, such as /f/ and /s/, the vocal cords stay open, but the articulators create a constriction in the vocal tract producing an air-flow turbulence, which is perceived as a speech sound. The articulators are involved in the formation of constriction, the manner and place of which determines the linguistic identity of the sounds. Since articulatory gestures are mainly responsible for producing different speech sounds (perceived as phones), they are often referred to as linguistic gestures.

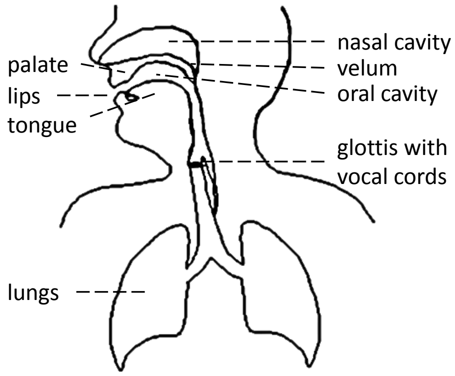


Figure 1: Human speech production physiology.

2.2.1 Acoustic theory of speech

The acoustic theory of speech (Fant, 1970) suggests a computational model to relate the speech production physiology with the acoustic properties of speech sounds. The computational model –also known as the source-filter model– represents the speech signal in terms of a convolution between a source excitation signal and a filter impulse response as shown in Figure 2. In the case of voiced sounds, the source excitation signal is a glottal pulse, whereas in the case of unvoiced sounds, the source excitation signal is the turbulence at the vocal tract constriction. The filter impulse response is the spectral characteristic of the resonance in the vocal tract. In speech, the vocal tract resonance frequencies are known as formants, and are indicative of both the linguistic content and the size and the shape of the vocal tract.

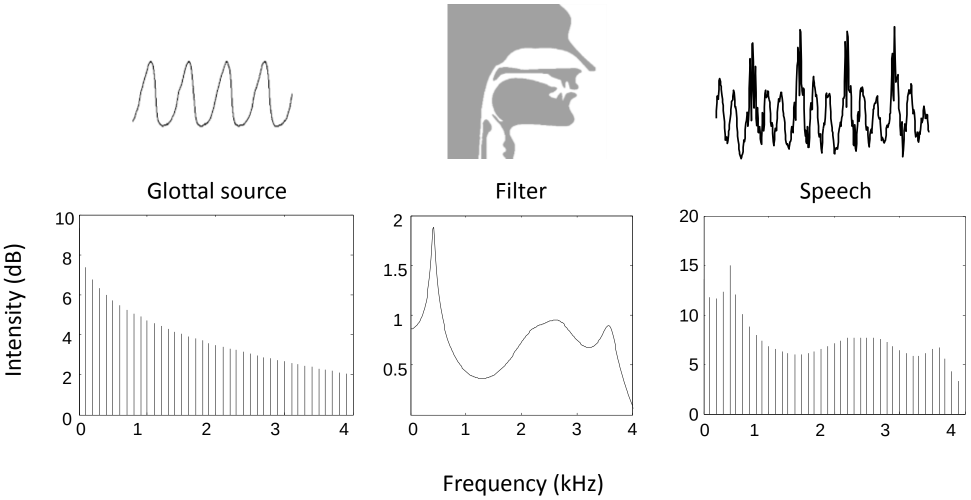


Figure 2: Acoustic theory of speech: speech signal (right) is the convolution of the glottal source excitation signal (left) and the vocal tract filter response (middle).

During speech production, the source and the filter characteristics keep changing over time. But, due to the quasi-stationary nature of speech, these characteristics can be treated as static for a small period of time (about 25 ms). Therefore, based on the acoustic theory of speech, we can model the speech production physiology as a slow-varying linear system (Figure 3). The time-varying linear system consists of: (i) source excitation signal generators (a pulse train for voiced sounds and a white noise generator for unvoiced sounds); (ii) a switch that selects the appropriate source signal based on voicing; and (iii) a vocal tract filter that modulates the excitation signal. This linear system allows us to analyze the speech signal in terms of the parameters that relate to speech production physiology, and to synthesize speech from those parameters.

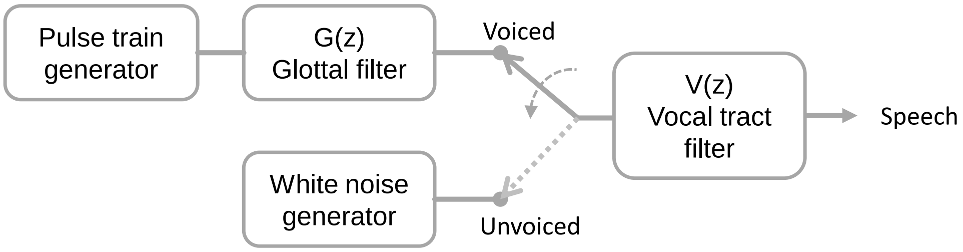


Figure 3: A simplified computational model of speech production physiology.

Computational models of speech such as the source-filter model provide a useful platform for speech modification. These models allow us to represent speech with meaningful theoretically-motivated model parameters that correspond to perceptual characteristics such as pitch, loudness and formants, and to synthesize speech from the modified parameters.

2.2.2 Speaker's identity, voice and accent

One of the evaluation criteria of the foreign accent conversion method is its ability to preserve the voice-quality of the non-native utterances. Although the voice-quality refers to the speaker's identity in many cases, in the context of foreign accent conversion, subtle differences between them needs to be emphasized. Several studies have shown that foreign accents, real or fake, deteriorate one's ability to identify a speaker (Tate, 1979; Torstensson *et al.*, 2004; Sjöström *et al.*, 2006; Sullivan and Schlichting, 2007). These studies show that the perceived identity of a speaker is not limited to the organic characteristics of the speaker; the speaker's identity is also associated with the linguistic gestures (e.g., accents) that arise from the cognitive and motor control of the articulators.

Because of the interaction between the speaker's identity and the accent, in the context of foreign accent conversion, it is important to differentiate the organic aspects (voice-quality) from the linguistic aspects (i.e. accent) of the speaker's identity. Thus, the objective of foreign accent conversion methods is to generate speech that matches the voice-quality of the non-native speaker, not necessarily the speaker's identity. To represent the voice-quality aspect of the speaker's identity, in this thesis, we use the term voice identity.

In the next section, we present a theory that explains speech signal as the interaction between these two components, voice-quality and the linguistic gestures.

2.2.3 Modulation theory

The source-filter model separates speech signal into two components, but it is not obvious what they represent perceptually. It can be assumed that the source signal represents speaker's identity while the filter represents the linguistic characteristics. However, Traunmüller (1985) observed that changing the formants alone can change the perception of voice quality. Similarly, changing the pitch alone can change the linguistic content of a sound. To accommodate for these observations, in the modulation theory of speech Traunmüller (2005) postulated, "A speech signal is basically the result of a process in which a carrier, characterized by the static properties of the speaker's voice, has been modulated by phono-articulatory gestures." In this view the 'carrier' or speaker's voice-quality is not only determined by the glottal source but also by the shape and size of the neutral vocal tract. Similarly, phono-articulatory gestures, which consist of pitch patterns and articulatory gestures, give linguistic color to the sound and are independent of voice-quality characteristics.

The modulation theory of speech is the motivation behind the development of articulatory-based foreign accent conversion methods in this dissertation work. It offers a principled way to transpose the voice-identity or the linguistic information across speakers via articulatory recordings. While the articulatory synthesizer embodies the anatomy and organic quality of voice, the articulatory trajectory represents the linguistic gestures in speech.

2.3 Speech analysis and synthesis

In this subsection, we briefly review the most common speech analysis and synthesis techniques used in speech modification. Among the four techniques reviewed here, the first three techniques are based on the source-filter model and its variants. The fourth one is based on harmonic plus noise model, which represents speech as the combination of a harmonic and a noise component separated in frequency domain.

2.3.1 Linear predictive analysis

In **linear predictive analysis** (Atal and Hanauer, 1971), the vocal tract response is modeled as an all-pole filter. The transfer function of the filter $V(z)$ is given as

$$V(z) = \frac{G}{1 - \sum_1^p \alpha_k z^{-k}}, \tag{1}$$

where p is the order of the model, G is the gain, and α_k are the linear prediction coefficients (LPC). The underlying assumption is that the sequence of values can be predicted as a linear combination of the finite number of preceding values. This model is widely used in speech analysis because estimating the model parameters and re-synthesis only requires linear filters. However, being an all-pole model, the linear predictive analysis is only appropriate to capture formants peaks, not the troughs (see Figure 4). Therefore, LPCs are not appropriate representation for nasals, which have a characteristic dip after the first formant in their spectrum. In addition to the misrepresentation, LPCs are also sensitivity towards small numerical errors, which significantly impacts the filter property (i.e. the pole locations), often leading to an unstable vocoding system. To improve stability, LPCs are generally converted into more stable representations such as linear spectral frequencies (LSF), specifically in speech modification problems (Arslan and Talkin, 1997; Kain and Macon, 1998).

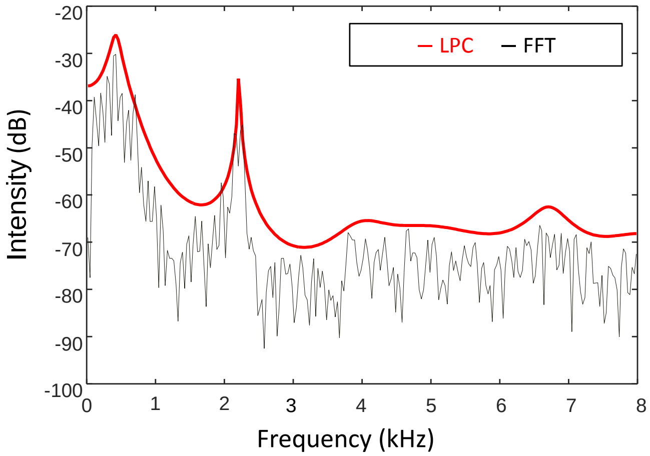


Figure 4: A typical FFT and LPC spectrum of a nasal speech segment.

2.3.2 Cepstral analysis

Cepstral analysis transforms the speech signal into cepstral coefficients, where the source excitation and the vocal tract impulse response become additive so that a linear separation of these components is possible. The cepstral coefficients are calculated by taking linear cosine transform of log power spectrum of speech. The source and the filter components can then be separated by liftering. The Mel frequency scale is commonly used to match the human auditory system, and the resulting cepstral coefficients are called Mel cepstral coefficients (MCCs). MCCs are the most common spectral representation used in speech synthesis as the Mel log spectrum approximation (MLSA) filter devised for Mel cepstral coefficients is shown to result in better quality speech than the LPC vocoder (Imai, 1983). Mel cepstral coefficients are also found suitable for statistical modeling as shown by their performance in statistical parametric synthesizer (Toda *et al.*, 2007; Zen *et al.*, 2007).

In another variant of cepstral analysis, the vocal tract spectra are represented as Mel frequency cepstral coefficients (MFCCs)(Davis and Mermelstein, 1980). MFCCs are calculated by passing the audio signal through Mel frequency filterbanks and taking DCT of the logarithms of output energies. Mel frequency filterbanks are the overlapping triangular filters uniformly spaced in Mel scale frequency inspired by human auditory perception. Due to the relation with human auditory system, MFCCs have become de-facto representation for speech recognition (Young, 1996).

It is also possible to mix cepstral analysis with other speech analysis techniques. For example, instead of using the power spectrum of speech signal, spectral envelope extracted from linear predictive analysis can also be used to estimate log spectral energies required to calculate cepstral coefficients. Such mixed analysis combines the advantages of both analysis techniques. For example, cepstral coefficients extracted from the LPC spectrum (known as LPC cepstrum) not only enjoy the robustness and naturalness of cepstral coefficients, but also the high intelligibility of LPC synthesis. In addition, LPCCs can be extracted more efficiently than MFCCs.

2.3.3 STRAIGHT analysis

STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) analysis uses bilinear interpolation over the time-frequency representation of the speech signal to estimate the smooth spectrogram (Kawahara, 1997). Illustrated in Figure 5, STRAIGHT analysis decomposes the speech signal into three independent components: (i) a smooth spectrogram free from the interference of fundamental frequency and the harmonics, (ii) a trajectory of fundamental frequency (f_0), and (iii) aperiodicity signal, which is the spectrogram of the nondeterministic excitation signal (e.g. noise). This model allows independent modification of these three components without any significant impact on the naturalness and the acoustic quality of the synthesis. Due to the naturalness of the synthesis and the flexibility of the model, STRAIGHT analysis and synthesis engine has gained popularity in applications such as voice conversion (Ohtani *et al.*, 2006; Toda *et al.*, 2007) and parametric text-to-speech synthesis (Zen *et al.*, 2013). For these same reasons, we use STRAIGHT as the speech analysis and synthesis platform in this dissertation work. To reduce the dimensionality of the STRAIGHT spectrum for statistical modeling, we extract MFCCs from the STRAIGHT spectrum.

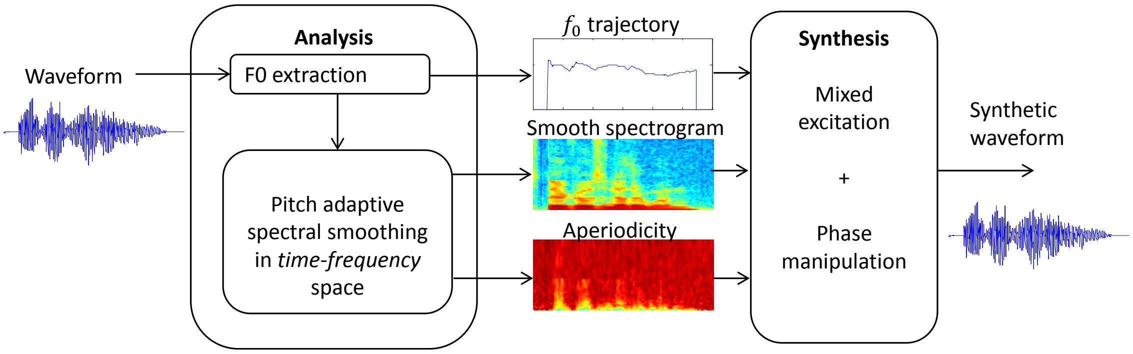


Figure 5: STRAIGHT analysis and synthesis.

2.3.4 Sinusoidal analysis and harmonics plus noise model

Sinusoidal analysis (McAulay and Quatieri, 1986) represents the speech signal as the sum of sinusoids as given by the equation (2), where L is the number of harmonics, f_0 is the fundamental frequency, and A_l and ϕ_l represent the amplitude and phase of the l^{th} harmonic, respectively. Representing speech as the sum of a fundamental frequency component and the harmonics preserves the naturalness of speech. Specifically, in the case of voiced sounds, the sinusoidal analysis gives more accurate representation of harmonics amplitudes than the LPCs and MFCCs. While MFCCs are known for smoothening of the vocal tract spectra, which leads to muffled speech, the LPCs are known for emphasizing the formants which makes the speech more robotic.

$$s(t) = \sum_{l=1}^{L} A_l \sin(2l\pi f_0 t + \phi_l) \tag{2}$$

However, the sum of sinusoids is not an appropriate representation for unvoiced sounds, and the voiced fricatives and affricates with high frequency noise. As a solution, Stylianou (2005), proposed a popular variation of sinusoidal model known as harmonic plus noise model (HNM). In HNM analysis model, spectra are divided in two frequency bands whose boundary can vary across frames. The lower band is represented as sinusoids at harmonic frequencies whereas the upper band is represented as whit-noise excited linear predictive model. In the case of unvoiced sounds, the boundary is at 0Hz. Such a hybrid representation is found to improve synthesis quality, especially, in pitch-time scaling (Stylianou, 2001) and voice-conversion (Stylianou *et al.*, 1998).

2.4 Articulatory speech processing

Articulatory methods capture the underlying mechanics of speech production by recording both the vocal tract anatomy and the kinematics of the articulators responsible for linguistic coloring of the speech signal. Because of the extra information, articulatory data allows more natural way to modify certain speech characteristics than the audio signal. For example, modifying vowel /i/ to /e/ only requires modifying the tongue height parameter in the articulatory data, a process that is more complex if done in the acoustic space. Similarly, articulatory data also provides the voice-quality independent representation of the linguistic content in speech. This property of the articulatory data is important in the context of foreign accent conversion because it allows transposing native linguistic gestures to the non-native speaker without altering the voice-quality.

Separating the voice-quality component from the linguistic content in the acoustic signal is known to be a challenging task (Felps *et al.*, 2009).

In this subsection, we review the speech processing techniques for speech modification in the articulatory domain. First, we discuss various articulatory representations and the common techniques to collect articulatory data during speech production. Secondly, we present the articulatory normalization techniques to account for the speaker-specific individual differences. Next, we review some of the data-driven articulatory synthesis methods as a vehicle to perform articulatory speech modification. Finally, we discuss articulatory inversion methods to estimate articulatory features from acoustic recordings. The inverted articulatory features allow us to use articulatory speech modification techniques without having to record the articulatory data, which is an expensive and invasive process.

2.4.1 Articulatory representation

Toutios and Margaritis (2003) classify articulatory representations into three groups. The first group is based on physical measurements such as electromagnetic articulography (EMA). The second group is based on theoretical models of speech production such as Maeda (Maeda, 1979), lossless tube and *tract variables* (TVs) (Browman and Goldstein, 1990). The third group is based on articulatory phonetics, a representation of speech in terms of abstract features such as manner and place of articulation, voicing, front-back, nasality, rounding, and stress. In what follows, first, we review the techniques for physical measurement of the articulatory configuration. Then, we discuss the different types of articulatory representations in the context of articulatory normalization and articulatory synthesis.

2.4.2 Articulatory measurements

Articulatory measurements provide two types of articulatory information: (i) the physical shape and size of the vocal tract and the articulators of a speaker, and (ii) the linguistic gestures. The shape and size of the vocal tract can be measured using magnetic resonance imaging (MRI). Such 3D images are useful in developing physical model of vocal tract (Birkholz *et al.*, 2006), but these images are highly redundant as the representative of the linguistic gestures. According to Ladefoged (1980), only sixteen articulatory parameters are necessary and sufficient to characterize all the possible sounds in all the known languages (see Table 1). Thus, the articulatory recording during speech production usually tracks the movement of certain flesh-points in the midsagittal plane of the vocal tract. For example, in the well-known XRMB dataset (Westbury, 1994) articulatory configurations were captured using x-ray microbeam that tracked the position of eight gold pellets attached to the different points of interest in the vocal tract. Similarly, the MOCHA-TIMIT corpus (Wrench, 1999) contains the coordinates of 8 flesh-points in the vocal tract captured using electromagnetic articulography (EMA) sampled at 200Hz.

Table 1: Sixteen articulatory parameters necessary and sufficient to characterize all possible sounds in human languages (Ladefoged, 1980).

i.	Front raising	ix.	Lip width
ii.	Back raising	x.	Lip Protrusion
iii.	Tip raising	xi.	Velic opening
iv.	Tip advancing	xii.	Larynx lowering
v.	Pharynx width	xiii.	Glottal aperture
vi.	Tongue bunching	xiv.	Phonation tension
vii.	Lateral tongue contraction	xv.	Glottal length
viii.	Lip height	xvi.	Lung Volume decrement

EMA has gained popularity for multiple reasons over other articulatory recording technologies. Compared to X-ray microbeam, EMA is safer and hence more suitable for collecting larger corpus. Compared to MRI, the high temporal resolution of EMA makes it more beneficial, specifically, to capture the dynamics of vocal tract configurations during continuous speech. In contrast, due to the low temporal resolution, MRI can miss some short lived articulatory configurations. Unlike EMA, MRI-based recordings also suffer from equipment noise.

However, the low spatial resolution of EMA data may not be sufficient to differentiate between a complete closure (in stops) and the formation of a small constriction (in fricatives). To overcome this problem, EMA recordings are often supplemented with additional measurement. Electropalatography (EPG) is commonly used in conjunction with EMA to record complementary information such as the location and duration of contacts between tongue and the hard palate.

In this dissertation work, we use EMA recordings as the articulatory representation. The corpus contains simultaneous recordings of EMA and audio signal from a native and a non-native speaker of American English (Felps *et al.*, 2012). The articulatory recording consists of six EMA pellet positions in the midsagittal cross section of the vocal tract. Illustrated in Figure 6, the EMA data are collected for the fleshpoints at the frontal oral cavity only —the EMA pellets cannot be placed at the back cavity because of the possible gag reflex. However, this limitation of EMA is not critical to this work. Since, according to the frontal cavity hypothesis (Hermansky and Broad, 1989) the frontal cavity is primarily responsible for linguistic coloring; “the back-cavity geometry is only a causal consequence and contributes mainly speaker-dependent information.”

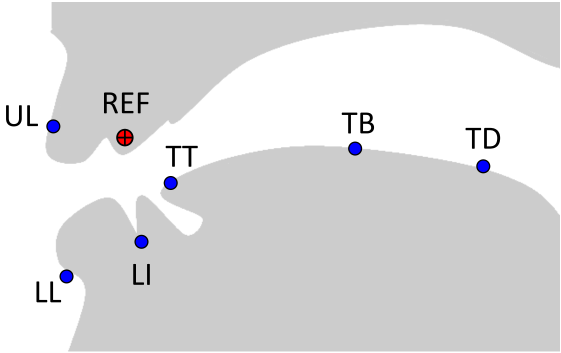


Figure 6: Position of the 6 EMA pellets used in our study; UL: upper lip; LL: lower lip; LI: lower incisor; TT: tongue tip; TB: tongue blade; TD: tongue dorsum. An additional pellet (red cross-hair) was placed on the upper incisor and served as a reference.

2.4.3 Articulatory normalization

Articulatory data needs to be normalized to remove the effect of anatomical differences in the vocal tract across speaker so that it can be applied across different speakers. One approach is to parameterize the measured articulatory positions into a speaker-independent representation. Several such representations have been suggested in the literature. As an example, Maeda (1990) proposed a set of relative measurements of the vocal tract that explain the majority of articulatory variance. In Maeda's representation, the vocal tract is represented by seven parameters: lips opening, jaw opening, lip protrusion, tongue tip height, tongue body shape, tongue dorsum position, and velum position. Al Bawab *et al.* (2008) developed a method to approximate Maeda parameters from EMA pellet positions; to remove individual differences, the method performed within-speaker z-score normalization of the approximated Maeda parameters. This normalized representation was then used for automatic speech recognition from articulatory positions derived from acoustics via analysis-by-synthesis. Hashi *et al.* (1998) proposed a normalization procedure to generate speaker-independent average articulatory postures for vowels. Using data from the X-ray microbeam corpus (Westbury, 1994), the authors scaled articulatory positions relative to a standard vocal tract, and then expressed the tongue surface relative to the palate. This procedure was able to reduce cross-speaker variance in the average vowel postures. *Tract variables* (TVs) (Browman and Goldstein, 1990) have also been used as speaker-independent articulatory representations. As an example, Ghosh and Narayanan (2011a) converted EMA articulatory positions into TVs, which were then used as the articulatory representation in a subject-independent articulatory inversion model. The authors reported inversion accuracies close to subject-dependent models, particularly for the lip aperture, tongue tip and tongue body articulators.

A second approach to account for individual differences is to learn a cross-speaker articulatory mapping. As an example, Geng and Mooshammer (2009) used the Procrustes transform, learned from a parallel corpus containing articulatory trajectories of multiple speakers during vowel production. The objective of the study was to unveil speaker-independent strategies for vowel production by removing speaker-specific variations. The authors reported a 30% improvement in subject-independent articulatory classification of vowels following Procrustes normalization. Qin *et al.* (2008) described a method to predict tongue contours (as measured via ultrasound imaging) from a few landmarks (EMA pellet positions). Using a radial basis function (RBF) network, the authors were able to reconstruct full tongue contours with 0.3-0.2mm errors using only 3-4 landmarks. In a follow-up study (Qin and Carreira-Perpinán, 2009), the authors proposed an articulatory mapping to adapt the previous predictive model to a new speaker using a 2D-wise linear alignment mapping. Their results show that a small adaptation corpus (about ten full tongue contours) is sufficient to recover very accurate (0.5 mm) predictive models for each new speaker. These studies suggest that a linear mapping can model a significant amount of inter-speaker differences in the vocal tract geometry.

More recently, Felps *et al.* (2014) extended the Procrustes transformation of EMA position data by allowing independent local translation at each articulatory fleshpoint and observed further reduction in the inter-speaker differences. The independent translation parameters for each fleshpoint allowed the transform to adjust for the non-uniform positioning of the articulatory fleshpoints across speakers.

Additional reduction in inter-speaker differences may be achieved by allowing independent scaling and rotation parameters for each fleshpoint.

2.4.4 Articulatory synthesis

Articulatory synthesizers have had a long tradition in speech research, starting with the electrical vocal tract analogue of Stevens *et al.* (1953). These models have improved our understanding of the speech production mechanism and in recent years have also provided alternative speech representations to improve the performance of automatic speech recognition systems (King *et al.*, 2007; Ghosh and Narayanan, 2011a; Arora and Livescu, 2013).

Articulatory synthesis methods can be grouped into two broad categories, physics-based models, and data-driven models. **Physics-based models** approximate vocal tract geometry using a stack of cylindrical tubes with different cross section areas. Speech waveforms are then generated by solving the wave propagation equation in the approximated tube model. In a classical study, Mermelstein (1973) analyzed midsagittal x-ray tracings to extract ten parameters that represented the configuration of the lips, jaw, tongue, velum and larynx. This parameterization was then geometrically converted into a vocal tract area function and the corresponding all-pole filter model, This study showed that the midsagittal position of a few critical articulators is sufficient to generate intelligible speech, and served as the basis for the articulatory synthesizer of Rubin *et al.* (1981). The midsagittal representation of articulators was also emphasized in another classical articulatory model by Maeda (1990). The author analyzed X-ray motion pictures of the vocal tract from two speakers to extract seven articulatory parameters, and found that 88% of the variance during phonetic articulation could be explained with only four articulatory parameters (three tongue points and jaw position). These early studies cemented the use of the vocal tract midsagittal plane as an articulatory representation in speech production research. Later research addressed the issue of generating articulatory trajectories from text using principles from articulatory phonology (Browman and Goldstein, 1990), leading to the development of the Task Dynamic Model (Saltzman and Munhall, 1989), and that of speech motor skill acquisition, resulting in the DIVA (Directions Into Velocities of Articulators) model (Guenther, 1994). A concern with articulatory synthesis models is the large number of parameters that need to be specified in order to produce an utterance, and the lack of guarantees that the resulting trajectories correspond to the actual articulatory gestures of a speaker. This makes it difficult to determine whether poor synthesis results are due to the generated articulatory gestures or the underlying articulatory-to-acoustic model. To address this issue, Toutios and Maeda (2012) coupled Maeda's model with articulatory positions measured from EMA and real-time magnetic resonance imaging (rtMRI). Visual alignment between EMA pellet positions, the standard Maeda vocal tract grid, and rtMRI was performed manually; from this, two geometrical mappings were computed: (i) a mapping from EMA to standard Maeda control parameters, and (ii) a mapping from the standard Maeda control parameters to a set of speaker-specific vocal tract grid variables. The authors were able to synthesize "*quite natural and intelligible*" VCV words; a subsequent study (Toutios and Narayanan, 2013) using the same procedure reported successful synthesis of French connected speech. However, voice similarity between the original speaker and the articulatory synthesis was not assessed as part of the study.

In contrast with physics-based models, **data-driven models** use machine learning techniques to build a *forward mapping* from simultaneous recordings of articulators and acoustics (Kaburagi and Honda, 1998; Toda *et al.*, 2008; Aryal and Gutierrez-Osuna, 2013). Because these models are generally trained on individual speakers, the resulting forward model automatically captures the voice characteristics of the speaker, making them ideally suited for accent conversion. In an early study, Kaburagi and Honda (1998) used a k-nearest-neighbors method to predict acoustic observations from articulatory positions. Given a target articulatory frame, estimating its (unknown) acoustic observation consisted of finding a few closest articulatory frames in the corpus, and then computing a weighted average of their acoustic observations. The authors found that synthesis quality improved when the search for the closest articulator frames was limited within phoneme category. In an influential study, Toda *et al.* (2008) proposed a statistical parametric approach to learn the forward mapping. The approach consisted of modeling the joint distribution of articulatory-acoustic vectors with a GMM. Given a target articulatory frame, its acoustic observation was estimated from the GMM using a maximum likelihood estimate (MLE) of the acoustic trajectory considering its dynamic.

Considering the dynamics of estimated acoustic features reduces unnatural spectral discontinuities across adjacent frames and improves the acoustic quality. But as

Nakamura *et al.* (2006) reported, the use of dynamic information not only at the output (acoustics) but also at the input (articulators) by modeling their trajectories using context-dependent hidden Markov models (HMM) increases the accuracy of articulatory-to-acoustic mapping. However, these improvements come at the expense of much higher computational costs during synthesis because of the iterative estimation process. Moreover, these methods also require the complete sequence of articulatory frames from a test utterance before their corresponding acoustics can be estimated –a further limitation for real-time synthesis applications. Thus, exploiting the temporal structure of speech without adversely impacting articulatory-synthesis time remains challenging. As a possible solution, we present a forward-mapping based on deep neural networks (DNN) for real-time articulatory synthesis in this work.

2.4.5 Articulatory inversion

Extraction of the articulatory configuration from the acoustic signal —known as articulatory inversion— is a well-studied hard problem in speech processing (Atal *et al.*, 1978; Richmond *et al.*, 2003; Livescu *et al.*, 2007; Qin and Carreira-Perpinán, 2007a; Ghosh and Narayanan, 2011b). Because of the expensive and invasive nature of the existing articulatory recording technologies, articulatory inversion offers a method to approximate the articulatory representations of speech without directly measuring them. The benefits of using inverted articulatory features have been established in several applications. For example, inverted articulatory features can be used to provide articulatory visual feedback in computer assisted pronunciation training (Youssef *et al.*, 2011). The inverted features have also been shown to boost speech recognition performance because of their relation with the speech production physiology, especially in noisy or pathological speech (Mitra *et al.*, 2010; Arora and Livescu, 2013). Moreover, the study of articulatory inversion has also enhanced our understanding of phonetics and phonology (Browman and Goldstein, 1992).

Articulatory inversion is a challenging problem. Given an articulatory configuration, acoustic signal can be generated by solving the wave propagation equations (Maeda, 1982; Birkholz and Jackel, 2003), but the inverse is not trivial. There is no analytical solution to the wave equations, and the problem is ill-posed, i.e. the same acoustic state can be the outcome of multiple articulatory configurations. Nonetheless, the physical constraints in the articulatory movement can be exploited to find unique solutions to the inverse problem (Qin and Carreira-Perpinán, 2007b). Several statistical models such as Gaussian mixture model (Toda *et al.*, 2008), canonical correlation analysis (Livescu and Stoehr, 2009), hidden Markov model (Hiroya and Honda, 2004), neural networks (Kello and Plaut, 2004), and deep neural networks (Uria *et al.*, 2012) have been explored to represent inverse mappings. Among them, the DNN-based inversion method is found to be the most accurate to date (Uria *et al.*, 2012), possibly because of the DNNs ability to exploit the temporal nature of speech by using contextualized input.

Most of the inversion methods mentioned above are speaker-dependent and require articulatory data from the speaker during the training phase. The one exception we are aware of is a study by Ghosh and Narayanan (2011a) that proposed a method for speaker-independent articulatory inversion using *tract variables* (TVs), a constriction-based relative measure of the articulatory configuration. The authors found that the inversion accuracy comparable to the speaker-specific inversion *and* that inverted features were effective in boosting speaker-independent recognition performance in noisy speech.

2.5 Summary

In this section, we discussed causes and consequences of foreign accents and reviewed several topics in acoustic and articulatory speech processing that are relevant to accent modification. We started with a review of speech production physiology in human and a simplified computational model known as the source-filter model. Computational models are useful because they represent speech in a parametric form suitable to modify specific speech characteristics. Then, we discussed the modulation theory of speech as the motivation behind our articulatory-based method for foreign accent conversion. The articulatory data captures the linguistic gestures independent of the voice-quality, while the complex interaction between linguistic gestures and voice-quality information makes it difficult to separate them in acoustic signal. Finally, we reviewed the existing methods to record the articulatory gestures and to synthesize speech using those gestures as the control parameters.

3. LITERATURE REVIEW[2]

This section reviews existing foreign accent conversion methods. Earlier work in accent modification was motivated by its application in spoken language conversion system, where the objective was to generate speech in language other than the one in which the speech corpus is available (Campbell, 1998). Later, the possible application of accent modification techniques in computer aided pronunciation training (Repp and Williams, 1987; Nagano and Ozawa, 1990; Probst *et al.*, 2002; Bissiri *et al.*, 2006) for non-native speakers motivated further research in this area. Because of the difficulty in modifying segmental characteristics, the focus was only on modifying the prosodic aspects of non-native accents in the earlier pronunciation training tools (Eskenazi, 1999). While the prosody is critical in parsing continuous speech (Celce-Murcia *et al.*, 1996), segmental errors are also equally responsible for degrading intelligibility (Rogers and Dalby, 1996). Thus, Derwing *et al.* (1998) suggested considering both segmental and supra-segmental (prosodic) features in pronunciation training. In this section, we review the existing methods for both the prosodic and segmental modifications of non-native accented speech. We also review prior work on evaluation of the foreign accent conversion methods.

[2] The review on the existing foreign accent conversion methods are reprinted with permission from "Reduction of non-native accents through statistical parametric articulatory synthesis," by Aryal and Gutierrez-Osuna, 2015. *J. Acoust. Soc. Am.*, 137, pp. 433-446. ©2015 Acoustical Society of America.

3.1 Foreign accent conversion

Foreign accent conversion is closely related to voice conversion but seeks a more elusive goal. In voice conversion, the objective is to convert an utterance from a source speaker to sound as if it had been produced by a different (but known) target speaker (Sundermann and Ney, 2003; Turk and Arslan, 2006). To do so, voice conversion techniques attempt to transform the two main dimensions of a speaker's voice individuality: physiological characteristics (e.g. voice quality, pitch range), and linguistic gestures (e.g. speaking style, accent, emotional state, etc.) Because the target speaker is known, evaluation of voice conversion results is relatively straightforward. In contrast, accent conversion seeks to combine the vocal tract physiology of a non-native learner (L2) with the linguistic gestures of a native teacher (L1). This is a far more challenging problem because it requires separating both sources of information; it also seeks to synthesize speech for which there is no ground truth –the L2 voice with a native accent, which also makes evaluation more challenging than in the case of voice conversion.

The existing foreign accent conversion methods can be grouped into two categories; the acoustic-based and the articulatory-based. While acoustic-based methods perform accent conversion in the acoustic domain, articulatory-based methods use articulatory data to transfer native accent to a non-native speaker.

3.2 Acoustic-based approach

Some aspects of accent are acoustically realized as prosodic features such as pitch trajectory, phoneme durations, and stress patterns. In these cases, a simple prosody modification alone can significantly reduce the perceived accent of an L2 utterance. As an example, modification of vowel durations can reduce the foreign accentedness in Spanish-accented English (Sidaras *et al.*, 2009) because there is a significant difference in vowel durations between both languages.

Modifying the prosody of an L2 utterance is straightforward because the target pitch and energy patterns and phoneme durations can be directly obtained from an L1 utterance of the same sentence. Once these prosodic features have been extracted, various techniques such as TD-PSOLA (Sundström, 1998; Yan *et al.*, 2004), FD-PSOLA (Felps *et al.*, 2009), and STRAIGHT (Aryal *et al.*, 2013) have been found effective in modifying prosodic cues to foreign accents. The phoneme durations of the L2 utterance can be matched with the reference L1 utterance by learning their ratio between the L1 and L2 speakers (Sundström, 1998; Felps *et al.*, 2009), or by force-aligning the L1 and L2 utterances using dynamic time warping (Aryal *et al.*, 2013). In the case of pitch, however, the L1 pitch trajectory needs adjustment to match the vocal range of the L2 speaker so that the identity of the L2 speaker is preserved. For this purpose, Sundström (1998) computed the mean pitch values of the L1 and L2 speaker. She used the quotient of these two values to scale the L1 pitch trajectory so that it matches the pitch range of L2 speaker. Felps *et al.* (2009) used a slightly different approach and shifted the L1 pitch trajectory (instead of scaling) to match the mean pitch value of the L2 speaker.

In most cases, though, prosodic modifications are not sufficient to achieve accent conversion. As an example, a few studies have shown that modification of phonetic realizations (i.e., segmental modification) is far more effective for accent reduction than prosody modification alone, both within regional accents of the same language (Yan *et al.*, 2004) and different languages (Felps *et al.*, 2009).

In early work, Yan *et al.* (2004) developed an accent-conversion method by exploiting differences in vowel formant trajectories for three major English accents (British, Australian, and General American). The authors learned a speaker-independent cross-accent mapping of formant trajectories by building a statistical model (a two-dimensional HMM) of vowel formant ratios from multiple speakers, and then extracting empirical rules to modify pitch patterns and vowel durations across the three regional accents. Once these 2D-HMMs and empirical rules had been learned from a corpus, the authors then adjusted the formant frequencies, pitch patterns and vowel durations of an utterance to match a target accent. In an ABX test, 78% of Australian-to-British accent conversions were perceived as having a British accent. Likewise, 71% of the British-to-American accent conversions were perceived to have an American accent. In both evaluations, changing prosody alone (pitch and duration pattern) led to noticeable changes in perceived accent, though not as significantly as incorporating formant modifications as well. The method hinged on being able to extract formant frequencies, so it cannot be easily extended to larger corpora because formant frequencies are ill-defined for unvoiced phones and cannot be tracked reliably even in voiced segments.

A segmental modification method for accent conversion suitable for both the voiced and unvoiced phones was proposed by Felps *et al.* (2009). The authors used SEEVOC (Paul, 1981) to split short-time spectra into a spectral envelope and a flat glottal spectrum. Then, they replaced the spectral envelope of an L2 utterance with a frequency-warped spectral envelope of a parallel L1 utterance and recombined it with the L2 glottal excitation; frequency warping was performed using a vocal tract length normalization function that matches the average formant frequencies of the two speakers (Sundermann and Ney, 2003). Modification of prosodic cues (phone duration and pitch contour) was performed via FD-PSOLA (Moulines and Charpentier, 1990). Listening tests showed a significant reduction in accent following segmental modification: when listeners were asked to rate accentedness in a 7-point Likert scale[3], accent-converted utterances were rated as being 'somewhat' accented (1.97 numeric rating) whereas original L2 utterances were rated as being 'quite a bit' accented (4.85 numeric rating). In contrast, prosodic modification did not achieve a significant reduction in accent (4.83 numeric rating). Listening tests of speaker identity with forward speech showed that segmental transformations (with or without prosodic transformation) were perceived as a third speaker, though the effect disappeared when participants were asked to discriminate reversed speech. The authors concluded that listeners used not only organic cues (voice quality) but also linguistic cues (accentedness) to discriminate speakers, which suggests that something is inevitably lost in the identity of a speaker when accent conversion is performed.

A few studies have attempted to blend L2 and L1 vocal tract spectra instead of completely replacing one with the other, as was done in (Felps *et al.*, 2009). In one such study, Huckvale and Yanagisawa (2007) reported improvements in intelligibility for Japanese utterances produced by an English test-to-speech (TTS) after blending their

[3] 1: Not at all, 3: Somewhat, 5: Quite a bit, 7: Extremely

spectral envelope with that of an utterance of the same sentence produced by a Japanese TTS. More recently, we presented a voice morphing strategy that can be used to generate a continuum of accent transformations between a foreign speaker and a native speaker (Aryal *et al.*, 2013). The approach performs a cepstral decomposition of speech into spectral slope and spectral detail as shown in Figure 7. Accent conversions are then generated by combining the spectral slope of the foreign speaker with a morph of the spectral detail of the native speaker. Spectral morphing is achieved by first representing the spectral detail through pulse density modulation and then averaging pulses in a pair-wise fashion (Shiga, 2009). This morphing technique provides a tradeoff between reducing the accent and preserving the voice identity of the L2 learner, and may serve as a behavioral shaping strategy in computer assisted pronunciation training.

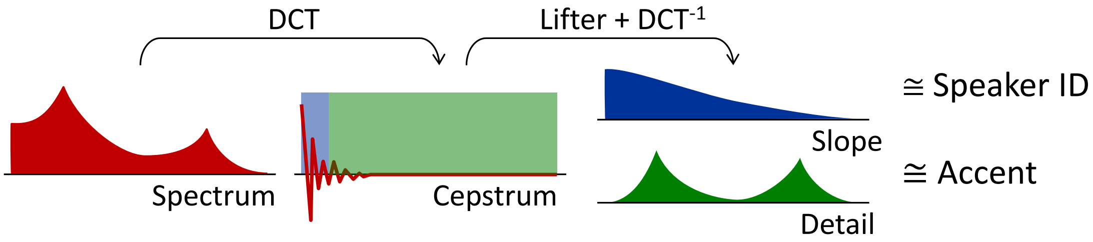


Figure 7: Cepstral decomposition of speech into spectral slope and spectral detail (DCT: Discrete cosine transform).

A limitation of both vocoding-based methods for accent conversion (Felps *et al.*, 2009; Aryal *et al.*, 2013) is that they require a careful alignment (at the frame level) of the parallel utterances from an L1 and L2 speaker. Given the common occurrence of deletion, substitution and insertion errors in L2 speech, however, obtaining a good alignment is not always possible. As mentioned earlier in the discussion of modulation theory of speech, the complex interaction of linguistic gestures and vocal tract physiology when looking at a spectrogram makes it difficult to separate them. As a result, accent conversions tend to be perceived as if they had been produced by a 'third-speaker,' one who is different from the original L1 and L2 speakers. Both of these issues disappear by operating in the articulatory domain. First, once an articulatory synthesizer has been built, there is no need for further alignment between L1 and L2 utterances: new accent conversions can be generated by driving the synthesizer directly from L1 articulators. Second, and more importantly, the linguistic gestures are readily available via the measured L1 articulators, whereas the voice identity is captured by the mapping from L2 articulators to L2 acoustics. Thus, in principle articulatory methods make it possible to achieve good accent conversion accuracy without compromising the voice identity of the L2 learner.

3.3 Articulatory-based approaches

The only prior work on articulatory accent conversion that we are aware of is a study by Felps *et al.* (2012) using unit-selection synthesis. Illustrated in Figure 8, the approach consisted of three stages, analysis, accent conversion and synthesis. During the analysis stage, the L1 and L2 utterances of the same sentences are segmented into diphone units. In the accent conversion stage, the mispronounced diphones in the L2 utterance are detected, and then replaced with other L2 diphone units (from L2 corpus). The replacing diphone units were selected such that they match the articulatory configurations of the corresponding diphone units in a reference native utterance. After segmental modifications during the accent conversion stage, the diphone units were concatenated during the synthesis stage and passed through the STRAIGHT engine to generate the audio waveform. During the synthesis phase, STRAIGHT is used to modify the pitch pattern of the concatenated speech to match the prosody of the reference L1 utterance.

The target articulatory feature vector consisted of six Maeda parameters (all but larynx height, which could not be measured with EMA), velocity for each of those parameters, normalized pitch, normalized loudness, and diphone duration. By replacing mispronounced diphones with other diphone units from the same speaker, this articulatory-based approach preserved the identity of the L2 speaker.

Unfortunately, the unit-selection synthesizer lacked the flexibility needed for accent conversion. First, the articulatory corpus contained 20,000 phones (or about 60 minutes of active speech) which, despite being larger than other articulatory databases (e.g., MOCHA-TIMIT (Wrench, 1999), X-Ray Microbeam (Westbury, 1994)), is considered small for unit-selection synthesis. Second, the unit-selection framework does not have a mechanism to interpolate between units, so it cannot produce sounds that have not been already produced by the L2 learner. Finally, the approach requires that L2 utterances be segmented and transcribed phonetically, which makes it impractical for pronunciation training settings. Based on these findings, we decided to explore other methods for articulatory synthesis that may have the flexibility and low-data requirements needed for accent conversion.

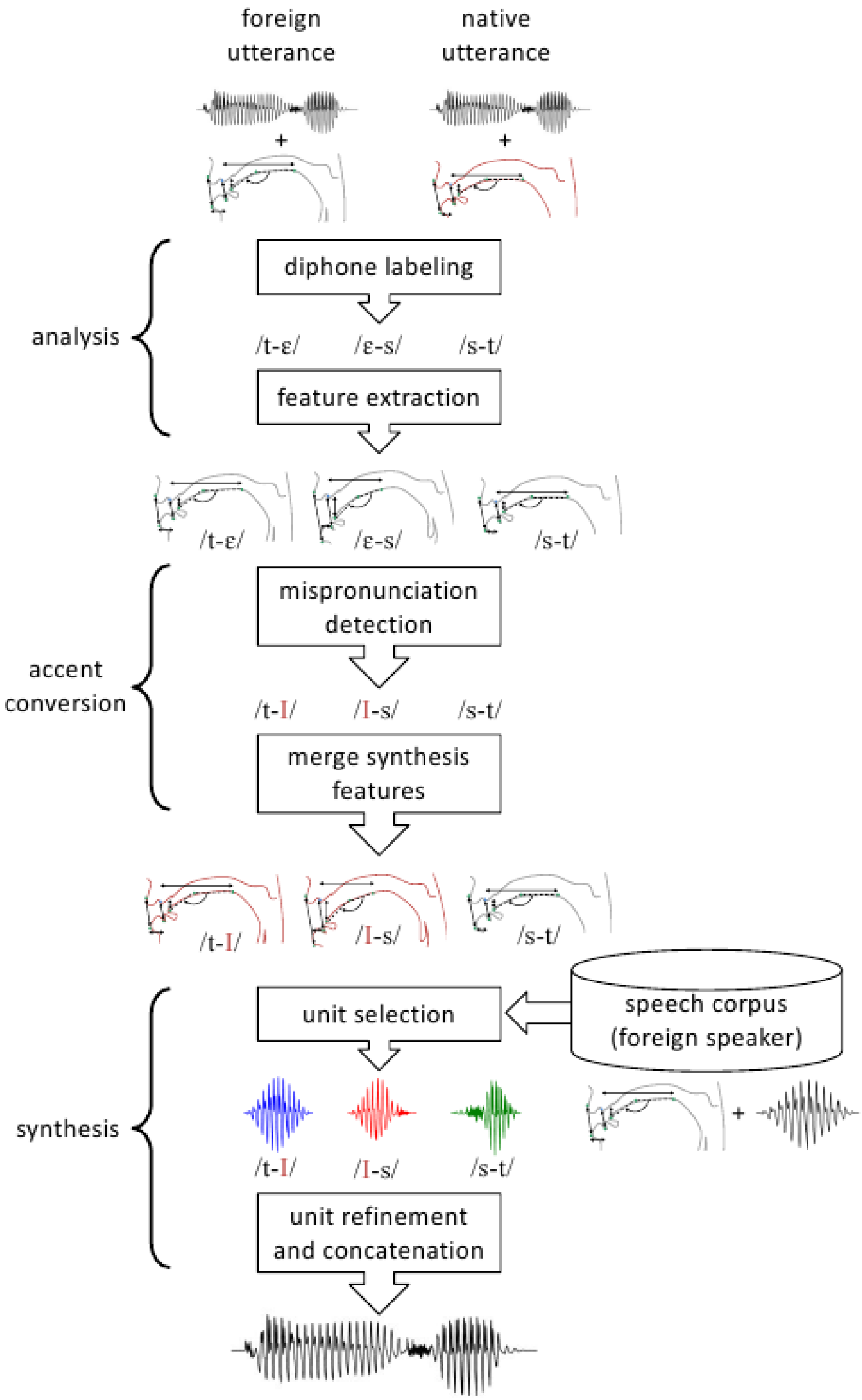


Figure 8: Articulatory foreign accent conversion based on unit selection (from Felps (2011)).

3.4 Evaluation of foreign accent conversion

The objective of foreign accent conversion methods is to generate native like utterances with the voice of a non-native learner. In addition, the resulting utterances are also required to be intelligible, natural and free from speech processing artifacts. Therefore, foreign accent conversions must be evaluated along three perceptual dimensions: acoustic quality, degree of non-native accents, and voice quality. In addition, intelligibility is also another important measure of foreign accent conversion, especially, in the evaluation of articulatory-based accent conversions. Because articulatory methods involve speech synthesis from articulatory data, which only provides partial information of the speech production apparatus that result in less intelligible synthesis (Kello and Plaut, 2004). In some cases, increased intelligibility has also been treated as the indicator of reduction in the perceived non-native accents (Huckvale and Yanagisawa, 2007). However, one must be careful when making such inferences since improved intelligibility may also be linked to the increased acoustic quality.

While the realistic evaluation of the accent conversion in all these perceptual dimensions requires subjective listening tests, several objective measures of these perceptual dimensions have been suggested for fast, low-cost and automatic assessment (Huckvale and Yanagisawa, 2007; Felps and Gutierrez-Osuna, 2010; Peabody, 2011). The objective measures, however, are restrictive in their application. For example, the objective measure based on ITU recommendation (ITU-T, 2004) was found highly correlated to the subjective quality ratings for a few speech coders in special test conditions, but in the quality assessment of foreign accent conversions (Felps and Gutierrez-Osuna, 2010) the measure was reliable only when averaged over 20 sentences, but not for evaluating the quality of individual utterance. Similarly, some objective measures (e.g., accent measure of Huckvale (2004)) are not applicable in foreign accent conversion because they are specific to a speaker, language or a set of words.

Since we use subjective evaluation of foreign accent conversion in this work for reliable measurements, in the following, we review various subjective perceptual evaluation methods in detail.

3.4.1 Acoustic quality assessment

The international telecommunication Union (ITU-T, 2006) recommends rating the utterances using mean opinion score (MOS) in a 5-point discrete scale (1:bad to 5:excellent). MOS is considered as a de-facto standard for subjective assessment of acoustic quality, and widely used in evaluation of the speech-modification techniques (Felps *et al.*, 2009). A more involved approach for quality assessment uses relative comparison between pairs of utterances. From a large set of pairwise similarity ratings between the utterances in the scale of 3 (much better) to -3 (much worse), a low dimensional embedding of the responses can be extracted that categorizes the utterances into groups that differ in quality ratings. One such low dimensional embedding can be extracted using Multi-dimensional scaling (Kruskal, 1964). This approach results in a more granular and reliable measurement of quality than MOS but also requires a large number of responses from the participants.

3.4.2 Intelligibility assessment

Intelligibility measures how accurately one can perceive the linguistic information in an utterance. One common measurement is the ratio of correctly identified words in the utterance calculated from the transcription (Lane, 1963; Barefoot *et al.*, 1993). Because the lexical context strongly influences the listener's ability to identify ambiguous sounds, the listener's familiarity with the language increases the word identification accuracy in a sentence level evaluation. Studies often use semantically unpredictable sentences (Pisoni and Hunnicutt, 1980; Goldstein, 1995) to account for such effect of linguistic structure in intelligibility. Similarly, due to learning effects, familiarity with a sentence (Davis *et al.*, 2005) also increases its intelligibility. Thus, when comparing multiple conditions using the same set of test sentences, we use different groups of listeners for each condition to account for the possible learning effect.

Subjective ratings have been used as a measure of intelligibility (Fayer and Krasinski, 1987; Munro and Derwing, 1995) to supplement word identification or transcription accuracy. These ratings provide an estimate of how confident the listeners are about the accuracy of the perceived linguistic information.

3.4.3 Assessment of non-native accents

The degree of non-native accents in an utterance can be evaluated using absolute rating in a scale spanning from a native to a reference non-native accented utterance (Munro and Derwing, 1995; Felps *et al.*, 2009; Felps *et al.*, 2012). The reference utterances are used to provide the anchor points for the listeners to calibrate their perceptual scale. However, it does not guarantee the consistency in their ratings over several sessions. In addition, the absolute ratings may not capture the subtle but perceivable differences in non-native accentedness between two utterances, because a high inter-rater variability can mask small differences in the accentedness. To detect such subtle differences, pairwise comparison of non-native accents tend to be more effective (Aryal *et al.*, 2013).

Who is a good judge of the non-native accents? Scovel (1988) found that the ability to gauge the degree of non-native accents develops in native speakers as they grow older, and among the non-native speakers with their exposure to the language. He also found that the adult native speakers have the best judgment of non-native accents. Therefore, adult monolingual native speakers of the language are the optimal participants for accent assessment tests. Furthermore, the known interaction between the acoustic quality of an utterance and its perceived non-native accentedness should also be considered in designing accent evaluation tests. By adding white noise to the utterances, Felps and Gutierrez-Osuna (2010) showed that utterances with lower acoustic quality are rated more non-native. Thus, when comparing the non-native accentedness in a pair of utterances, it is important to keep their acoustic quality comparable to avoid the quality bias in the perception of non-native accentedness.

3.4.4 Assessment of voice-identity

In voice conversion, ABX tests are commonly used to evaluate how close the voice conversion is to the target utterance compared to the source utterance (Kain and Macon, 1998; Toda *et al.*, 2007; Toth and Black, 2007). ABX test has also been used in accent conversion (Felps *et al.*, 2009). But, in order to reduce the effect of accents in perceived identity, Felps *et al.* (2009) played the utterances backward. The backward speech has recognizable timbre, variability on pitch (Black, 1973) but the prosodic and segmental information related to the accent is largely inaccessible to the listener (Munro *et al.*, 2010).

However, as Kain and Macon (2001) pointed out, the ABX test is not a true test of voice-similarity but the test of relative closeness to the target speaker as opposed to the source speaker. Instead, a same/different test is a more realistic test. In one such voice-similarity assessment, Kreiman and Papcun (1991) had participants listen to a pair of utterances and then asked them if the pair was from the same speaker or not and rate their confidence in a 7-point empirically grounded, well anchored (EWGA) scale. This measure provides the voice-similarity between the two speakers independent of any reference speaker.

4. ACOUSTIC-BASED FOREIGN ACCENT CONVERSION USING VOICE CONVERSION[4]

The existing accent conversion methods in acoustic domain follow a direct approach where a non-native utterance (L2) is modified such that it matches the prosodic and segmental characteristics of a reference native (L1) utterance (Campbell, 1998; Huckvale and Yanagisawa, 2007; Yan *et al.*, 2007; Felps *et al.*, 2009; Aryal *et al.*, 2013). These vocoding techniques attempt to decompose the spectral envelopes from a native (L1) and a non-native (L2) utterance into the components responsible for the linguistic gestures (e.g. accents) and the voice-identity components, then transpose these components across speaker.

In this section, we present an acoustic-based foreign accent conversion method that uses cross-speaker spectral mappings to estimate the trajectories of equivalent L2 acoustic features from a given sequence of L1 acoustic features from a reference native utterance as in voice conversion. By using cross-speaker spectral mappings, we not only avoid the difficulty of separating the linguistic and voice-quality related information from the spectral envelopes, but also obviate the error-prone time-alignment between the L1 and L2 utterances during conversion, the two main limiting factors in vocoding-based techniques. Unlike voice conversion, foreign accent conversion seeks to preserve the speaking style (accents) of the source speaker; therefore, the mappings learned for voice

[4] The description of the method and the experimental results are reprinted with permission from "Can voice conversion be used to reduce non-native accents?" by Aryal and Gutierrez-Osuna, 2014. *Proceedings of ICASSP*, pp. 7929-7933, ©2014 IEEE.

conversion cannot be directly applied to the accent conversion. This section shows how the cross-speaker statistical mappings of voice conversion can be adjusted during the training phase so that the trained mappings can be used for foreign accent conversion.

4.1 Foreign accent conversion based on spectral mapping

As shown in Figure 9, the proposed accent conversion method consists of two phases: training and conversion. During the training phase, we first use STRAIGHT use STRAIGHT (Kawahara, 1997) to extract the spectral features (MFCCs) and fundamental frequency (f_0) for the parallel training utterances from both the L1 and L2 speakers. After segmenting the utterances into frames, we pair each L1 frame with the acoustically closest L2 frames and vice versa. Then, using the frame pairs, we train a Gaussian mixture model (GMM) on the joint distribution of the spectral feature vectors from L1 and L2. Finally, we calculate means and standard deviations of $\log(f_0)$ for both the speakers and build a pitch modification (PM) function.

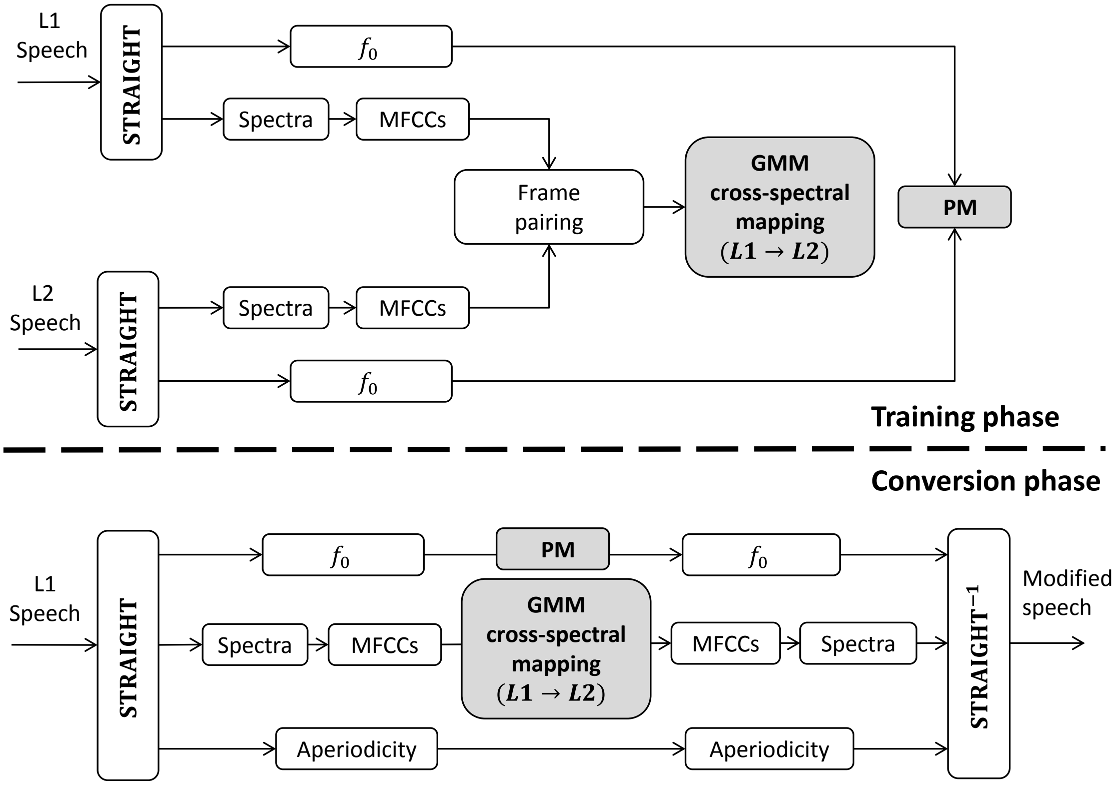


Figure 9: Foreign accent conversion method using cross-speaker statistical mappings.

During the conversion phase, we pass a test L1 utterance through the same feature extraction module as in the training phase. Once the pitch and spectral features are extracted, we use the pitch modification module and the trained cross-speaker spectral mappings to estimate the equivalent pitch trajectory and the spectral features for the L2 speaker, respectively. Given these modified parameters, STRAIGHT, finally, generates audio signal. More details on STRAIGHT feature extraction and synthesis, pairing of the L1 and L2 frames, pitch modification, and GMM-based mapping are given below.

4.1.1 STRAIGHT feature extraction and synthesis

We use STRAIGHT to extract acoustic features and synthesize the resulting speech waveform. Given an utterance, we extract f_0, *aperiodicity* and *spectral envelope* with STRAIGHT. For each frame (sampled at every 5ms in this study), we then compute $MFCC_{0-24}$ by warping the STRAIGHT spectral envelope according to the Mel frequency scale (25 Mel filterbanks, 8 kHz cutoff frequency) and applying a type-II discrete cosine transformation (DCT).

During synthesis, we reconstruct the STRAIGHT *spectral envelope* from the estimated spectral coefficients ($MFCC_{0\text{-}24}$). Specifically, given a vector of predicted MFCCs, the least-squares estimate of the spectral envelope is $\hat{\boldsymbol{s}} = (\boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^{\boldsymbol{T}}\boldsymbol{e}$, where $\boldsymbol{F}$ is the Mel Frequency Filter Bank (MFB) matrix used to extract MFCCs from the STRAIGHT spectrum, and $\boldsymbol{e}$ is the exponential of the inverse DCT of MFCCs. In a final step, we use the STRAIGHT synthesis engine to generate the waveform using the estimated *spectral envelope* $\hat{\boldsymbol{s}}$, the *aperiodicity* and the modified f_0.

4.1.2 Pairing acoustic vectors

In conventional voice conversion the source and target acoustic vectors are paired using forced alignment in a parallel corpus (Abe *et al.*, 1988; Toda *et al.*, 2007). Because of the systematic nature of the accent-related deviations, the mapping learned using the time-aligned parallel corpus is also likely to learn the accents of the non-native speaker. As a solution, our approach consists of pairing source and target vectors based on their acoustic similarity following vocal tract length normalization (VTLN). Both pairing approaches are illustrated in Figure 10.

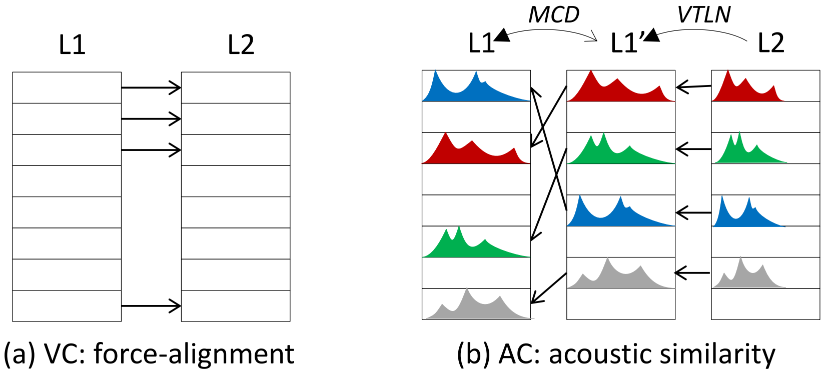


Figure 10: (a) Conventional approach to voice conversion; source and target utterances are paired based on their ordering in a forced-aligned parallel corpus, (b) Our approach to accent conversion: source and target utterances are paired based on their acoustic similarity following vocal-tract-length normalization (VTLN), MCD: Mel Cepstral Distortion.

The first step in our acoustic-similarity based pairing is to apply VTLN in order to reduce physiological differences in the vocal tract of the two speakers[5]. For this purpose, we use dynamic time warping to align parallel utterances from the L1 and L2 speakers, each utterance represented as a sequence of 24 Mel Frequency Cepstral Coefficients (MFCCs). Following Panchapagesan and Alwan Panchapagesan and Alwan (2009), we then learn a linear transform between the MFCCs of both speakers using least squares:

[5] More elaborate forms of speaker normalization may be used, such as context-dependent VTLN (Maragakis and Potamianos, 2008) or even speaker adaptation techniques, but this increases the risk of capturing not only physiological differences but also accent.

$$W = \arg\min \|\boldsymbol{u} - W\boldsymbol{y}_{\boldsymbol{L1}}\|^2 \quad (3)$$

where $\boldsymbol{u}$ and $\boldsymbol{y}$ are vectors of MFCCs from the L1 and L2 speakers, respectively, and $W$is the VTLN transform. Next, for each L1 vector $\boldsymbol{u_i}$ we find its closest L2 vector $\boldsymbol{y}_{\boldsymbol{j}}^{*}$ as:

$$\boldsymbol{y}_{\boldsymbol{j}}^{*} = \arg\min_{\forall \boldsymbol{y}} \|\boldsymbol{u_i} - W\boldsymbol{y}\|^2 \quad (4)$$

To make the search for the closest frame more efficient, we first group all L2 acoustic frames into 512 clusters using *k*-means. Then, for each L1 frame $\boldsymbol{u_i}$, we first find the closest L2 cluster and then the closest frame from those within that cluster. We repeat the process for each L2 vector $\boldsymbol{y_i}$ to find its closest match $\boldsymbol{u}_{\boldsymbol{j}}^{*}$:

$$\boldsymbol{u}_{\boldsymbol{j}}^{*} = \arg\min_{\forall \boldsymbol{x}} \|\boldsymbol{u} - W\boldsymbol{y}_i\|^2 \quad (5)$$

This results in a lookup table where each L1 and L2 vector in the database is paired with the closest vector from the other speaker. It is this lookup table that we then use to train a GMM, as explained next.

4.1.3 Cross-speaker spectral mapping

The cross-speaker spectral mapping is adopted from Toda et al. (Toda *et al.*, 2007), which uses we a GMM-based method for maximum likelihood estimation of spectral parameter trajectories considering the global variance of the target speaker[6]. Let $\boldsymbol{U_t} = [\boldsymbol{u}_t, \boldsymbol{\Delta u}_t]$ be a vector of static and dynamic (delta) MFCCs for the L1 speaker at

[6] For this study, we used our own MATLAB implementation of the GMM method of Toda *et al.* (2007).

frame $\boldsymbol{t}$, and $\boldsymbol{Y_t} = [\boldsymbol{y}_t, \Delta\boldsymbol{y}_t]$ be the corresponding vector for the L2 speaker. Then, we model the joint distribution $\boldsymbol{Z}_t = [\boldsymbol{U}_t, \boldsymbol{Y}_t]$ as

$$P\left(\boldsymbol{Z}_t \middle| \boldsymbol{\lambda}^{(z)}\right) = \sum_{m=1}^{M} \alpha_m \,\mathcal{N}(\boldsymbol{Z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}), \tag{6}$$

where $\boldsymbol{\lambda}^{(z)} = \left\{\alpha_m, \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}\right\}$ are the GMM parameters (weight, mean and covariance of the m^{th} mixture component, respectively), learned from a training set of joint vectors $\boldsymbol{z}_t$ using expectation-maximization (EM).

Given a trained GMM, we calculate the maximum likelihood estimate of acoustic features considering the dynamics and the global variance (GV) as follows. Let $\boldsymbol{U} = [\boldsymbol{U_1}, \boldsymbol{U_2} \dots \boldsymbol{U_T}]$ denote the sequence of L1 acoustic vectors in a source sentence. Consider also the within-sentence variance of the d^{th} acoustic feature $y_t(d)$ given by $v(d) = E[(y_t(d) - E[y_t(d)])^2]$. Thus, the GV of the static acoustic feature is written as $\boldsymbol{v}(\boldsymbol{y}) = [\,v(1), v(2) \dots v(D)]$ where D is the dimension of $\boldsymbol{y}_t$, and $\boldsymbol{y}$ is the sequence $[\boldsymbol{y}_1, \boldsymbol{y}_2 \dots \boldsymbol{y}_T]$. Now, the time sequence of estimated acoustic vectors (static only) is given by:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}} \; P\,(\boldsymbol{Y}|\boldsymbol{U}, \boldsymbol{\lambda}^{(z)}, \boldsymbol{\lambda}^{(v)}) \tag{7}$$

where $\boldsymbol{\lambda}^{(v)} = \{\boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(vv)}\}$, $\boldsymbol{\mu}^{(v)}$ is the vector of average variance for all acoustic features and $\boldsymbol{\Sigma}^{(vv)}$ is the corresponding covariance matrix, learned from the distribution of $\boldsymbol{v}(\boldsymbol{y})$ in the training set. The likelihood $P(\boldsymbol{Y}|\boldsymbol{U}, \boldsymbol{\lambda}^{(z)}, \boldsymbol{\lambda}^{(v)})$ is computed as

$$P\left(\boldsymbol{Y}\middle|\boldsymbol{U}, \boldsymbol{\lambda}^{(z)}, \boldsymbol{\lambda}^{(v)}\right) = P\left(\boldsymbol{Y}\middle|\boldsymbol{U}, \boldsymbol{\lambda}^{(z)}\right)^{w}.\; P\left(\boldsymbol{v}(\boldsymbol{y})\middle|\boldsymbol{\lambda}^{(v)}\right) \tag{8}$$

The distribution of GV, $P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)})$, is modeled by a single Gaussian $\mathcal{N}(\boldsymbol{v}(\boldsymbol{y}); \boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(vv)})$. The power term w $(= 1/2T)$ in equation (8) controls the balance between the two likelihoods.

Following Toda *et al.* (2007), we solve for $\widehat{\boldsymbol{y}}$ in equation (7)(23) iteratively via Expectation-Maximization. Namely, we define the auxiliary function with respect to $\widehat{\boldsymbol{y}}$ as:

$$Q(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}) = w \sum_{m} P(m|\boldsymbol{u}, \boldsymbol{Y}, \boldsymbol{\lambda}^{(z)}) \log P(\widehat{\boldsymbol{Y}}, m|\boldsymbol{U}, \boldsymbol{\lambda}^{(z)}) + \log P(\boldsymbol{v}(\widehat{\boldsymbol{y}})|\boldsymbol{\lambda}^{(v)}) \qquad (9)$$

At each M-step, we iteratively update the estimate of the trajectory (static elements only) as:

$$\widehat{\boldsymbol{y}} \leftarrow \widehat{\boldsymbol{y}} + \alpha\, \boldsymbol{\Delta}\widehat{\boldsymbol{y}} \qquad (10)$$

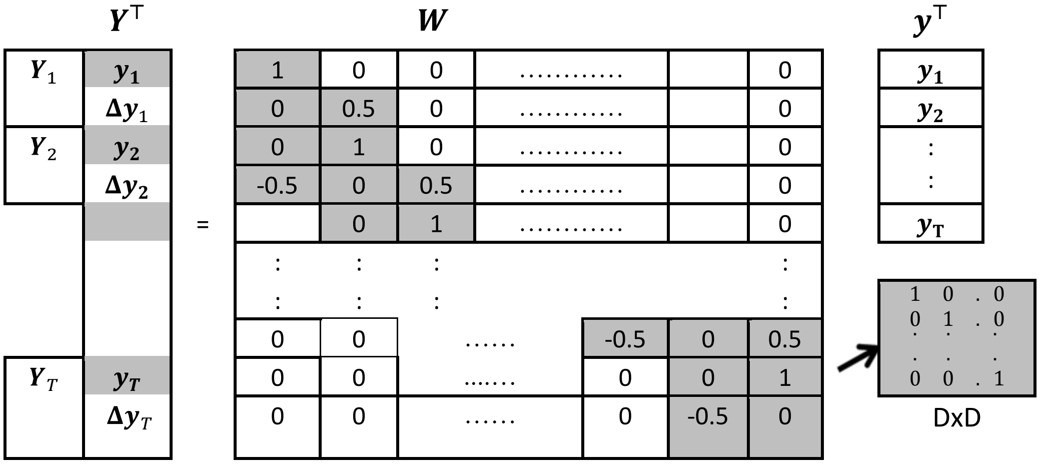
where α is a step-size parameter, and the steepest-descent gradient $\boldsymbol{\Delta}\widehat{\boldsymbol{y}}$ is given by

$$\boldsymbol{\Delta}\widehat{\boldsymbol{y}} = \delta Q(\boldsymbol{Y}, \widehat{\boldsymbol{Y}})/\delta\widehat{\boldsymbol{y}} = \omega\,(-\widehat{\boldsymbol{y}}\boldsymbol{W}^{\top}\overline{\boldsymbol{D}}\,\boldsymbol{W} + \overline{\boldsymbol{\Psi}}\,\boldsymbol{W}) + [\mathbf{v}_1, \mathbf{v}_2 \dots \mathbf{v}_t \dots \mathbf{v}_T] \qquad (11)$$

Vector $\mathbf{v}_{\mathrm{t}}$ in equation (11) is the GV-related adjustment of the target acoustic features at frame t, the dth element of which is computed as:

$$\mathrm{v}_t(d) = -\frac{2}{T}\,\boldsymbol{p}_{\boldsymbol{v}}(d)(\boldsymbol{v}(\widehat{\boldsymbol{y}}) - \boldsymbol{\mu}_{\boldsymbol{v}})(\,\hat{y}_t(d) - \sum_{t=1}^{T} \hat{y}_t(d)) \qquad (12)$$

where $\boldsymbol{p}_{\boldsymbol{v}}(d)$ is the dth column of $\boldsymbol{\Sigma}^{(vv)^{-1}}$. $\boldsymbol{W}$ is the *2DT*×*DT* matrix that translates a trajectory of the static parameters to a trajectory of the complete acoustic feature vector as given by:



(13)

In equation (11)**,** $\overline{\boldsymbol{D}}$ is a block-diagonal matrix whose diagonal consists of T covariance sub-matrices $\sum_{m=1}^{M} P(m|\boldsymbol{u}_t, \boldsymbol{Y}_t, \boldsymbol{\lambda}^{(v)})\, \boldsymbol{D}_m^{(Y)^{-1}}$; $t = 1 \dots T$ and $\overline{\boldsymbol{\Psi}}$ is a row vector of length $2DT$ consisting of T sub-vectors of length $2D$ given by $\sum_{m=1}^{M} P(m|\boldsymbol{u}_t, \boldsymbol{Y}_t, \boldsymbol{\lambda}^{(v)})\, \boldsymbol{D}_m^{(Y)^{-1}} \boldsymbol{E}_{m,t}^{(Y)}$; $t = 1 \dots T$, where

$$\boldsymbol{D}_m^{(\boldsymbol{Y})} = \boldsymbol{\Sigma}_m^{(YY)} - \boldsymbol{\Sigma}_m^{(Yu)} \boldsymbol{\Sigma}_m^{(uu)^{-1}} \boldsymbol{\Sigma}_m^{(uY)} \tag{14}$$

and $\boldsymbol{E}_{m,t}^{(Y)}$ is the conditional expected value as given by

$$\boldsymbol{E}_{m,t}^{(Y)} = \boldsymbol{\mu}_m^{(Y)} + (\boldsymbol{u}_t - \boldsymbol{\mu}_m^{(u)}) \boldsymbol{\Sigma}_m^{uu^{-1}} \boldsymbol{\Sigma}_m^{(uY)} \tag{15}$$

The algorithm requires an initial estimate of the trajectory of the static acoustic features $\widehat{\boldsymbol{y}}$. In our implementation, we initialize with the minimum mean square error (MMSE) estimate, which ignores the dynamics and global variance of the acoustic features. Given a trained GMM $\{\alpha_m, \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}\}$, the MMSE estimate is calculated by summing the conditional expected values from all Gaussian mixture components, weighted by their conditional membership probability for the given articulatory feature vector $\boldsymbol{u}_t$:

$$\hat{\boldsymbol{y}}_{t,init} = \sum_{m=1}^{M} P\big(m \big| \boldsymbol{u}_t, \boldsymbol{\lambda}^{(z)}\big) \boldsymbol{E}_{m,t}^{(y)} \tag{16}$$

where $\boldsymbol{E}_{m,t}^{(y)}$ is the static-feature-only subset of $\boldsymbol{E}_{m,t}^{(Y)}$, as given by equation (15).

4.1.4 Prosody modification

Following Toda *et al.* (2007), we use the aperiodicity and pitch trajectory of the source (L1) speaker, which captures the native intonation pattern, but normalize it to the pitch range of the target (L2) speaker to preserve his or her natural vocal range (see Figure 11). More specifically, given an L1 pitch trajectory $f_1(t)$, we follow the methods commonly used in voice conversion (Stylianou *et al.*, 1998; Toda *et al.*, 2007), and generate the modified L2 pitch trajectory $f_2(t)$ as:

$$\log\big(f_2(t)\big) = [\log(f_1(t)) - \mu_1] \frac{\sigma_2}{\sigma_1} + \mu_2 \tag{17}$$

where (μ_1, σ_1) and (μ_2, σ_2) are the mean and standard deviation of log-scaled pitch of the L1 and L2 speakers, respectively, calculated from the training corpus. This approach has two main advantages. First, it accounts for the dynamic range (not only the mean value) of the speaker. Secondly, it manipulates pitch in logarithmic scale similar to the human auditory system.

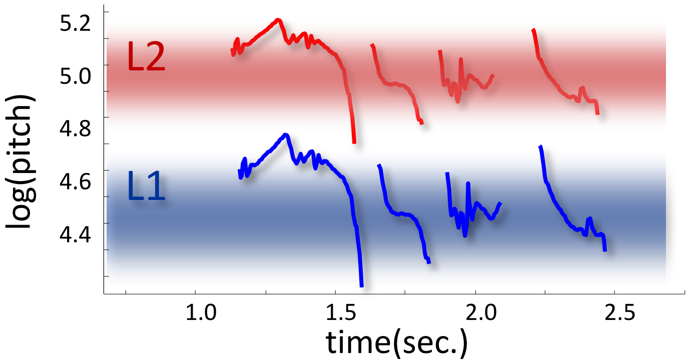


Figure 11: Shifting and scaling L1 pitch trajectory to match the vocal range of the L2 speaker.

4.2 Experimental

4.2.1 Conversion from non-native to native accent

To test the effectiveness of our accent conversion (AC) model, we compared it against the utterances from the L1 and L2 speakers. To evaluate the effect of acoustic-similarity based frame pairing, we also compared AC against the conventional voice conversion (VC). The baseline VC model was similar to the AC model except that the GMM model was trained on DTW-aligned pairs from source and target speakers –see Figure 10a, whereas the AC model was trained on acoustically-matched pairs as described in section 4.1.2. In both cases, the GMM consisted of 128 Gaussian components.

We performed two sets of perceptual listening tests:

1. *Perceived accent*: subjects listened to pairs of utterances (AC-VC, AC-L2, VC-L2) and were asked to select the utterance that sounded the least accented. Order of presentation in the pairs was randomized within subjects.
2. *Perceived speaker identity*: subjects listened to three utterances (A,B,X) and were asked to select whether the speaker in utterance X sounded closer to the identity in A or B. Utterances in X were AC; utterances A, B were either L1 or L2 (order of presentation was randomized within subjects). Following (Felps *et al.,* 2009), utterances were played *backward* to avoid interactions between accent and identity.

To ensure that the loss of quality in the AC and VC methods due to the MFCC compression step did not affect the perceptual ratings, control utterances from the L1 and L2 speaker were compressed to MFCC and then resynthesized as described in (Aryal and Gutierrez-Osuna, 2013).

4.2.2 Conversion from native to non-native accent

We also tested the effectiveness of our AC method to map accents in the opposite direction, i.e., imparting a non-native accent to the voice of a native speaker. For this purpose, we trained AC and VC models in a manner similar as in section 4.2.1, except we used L2 as the source speaker, and L1 as the target speaker. The six types of synthesis evaluated are summarized in Table 2.

Table 2: Summary of the six synthesis models[7] (AP: aperiodicity from STRAIGHT; DTW: dynamic time warping; AC: accent conversion; VC: voice conversion; '12' denotes transformation from L1 to L2).

Synthesis model	Frame pairing	Source MFCC	Target MFCC	AP
AC12	Acoustic	L1	L2	L1
VC12	DTW	L1	L2	L1
AC21	Acoustic	L2	L1	L2
VC21	DTW	L2	L1	L2
L1	-	L1	L1	L1
L2	-	L2	L2	L2

4.2.3 Experimental corpus

The speech corpus consisted of parallel recordings from a non-native speaker (whose first language was Spanish) and a native speaker of American English, previously described in (Felps *et al.*, 2012). Both subjects recorded the same 344 sentences chosen from the Glasgow Herald corpus. In addition, the non-native speaker recorded 305 sentences not spoken by native speaker. Out of the 344 sentences shared among both speakers, we randomly selected 294 sentences to train the GMM, and saved the remaining 50 sentences for testing purposes. For each sentence, we computed 25 MFCCs ($MFCC_{0:}$ energy; $MFCC_{1\text{-}24:}$ spectral envelope) as well as *pitch* and *aperiodicity* from the STRAIGHT (Kawahara, 1997) spectrum sampled at interval of 5ms[8].

[7] Audio samples are available in http://psi.cse.tamu.edu/samples/acvc.html

[8] STRAIGHT was also used to resynthesize utterances from the output of the GMM-GV model.

4.3 Results

Listening tests were performed on Amazon's Mechanical Turk. Following (Felps *et al.*, 2012), in order to qualify for the experiments participants first had to pass a screening test that consisted of identifying various American English accents: Northeast (i.e. Boston, New York), Southern (i.e. Georgia, Texas, Louisiana), and General American (i.e. Indiana, Iowa). Participants were also asked to list their native language/dialect and any other fluent languages that they spoke. If a subject was not a monolingual speaker of American English then their responses were excluded from the results. In the quality and accent evaluation tests, participants were asked to transcribe the utterances to ensure they paid attention to the recordings. Participants with incomplete responses were excluded from the study.

4.3.1 L1→L2 accent/voice conversion

Twenty participants rated the accent and identity of the AC12 and VC12 models on a set of 12 sentences, randomly selected from the 50 sentences in the test set. Both models were perceived to be less foreign-accented than the original L2 utterances. On average, listeners found VC12 to be less accented than the original L2 utterances 90% of the times (std. 9%). Likewise, listeners found AC12 less foreign-accented than L2 89% of the times (std. 9%). This result would suggest that there is no significant difference in accent reduction between conventional voice conversions (VC12) and our proposed method (AC12). However, when both models were compared against each other, participants found AC12 to be less accented than VC12 60% of the times (std. 10%).

The difference in perceived accent between the two models was statistically significant ($t = 4.19$, $p < 0.001$, single-tail).

Results from the ABX identity test show that participants found AC12 closer to L2 than to L1 an average of 77% of the times (std. 22%), which is statistically significant ($t = 5.51$, $p < 0.001$, single-tail) compared to chance levels (50%).

In summary, these results indicate that the proposed AC method is more effective in reducing accent than conventional VC, while at the same time it preserves the identity of the L2 speaker.

4.3.2 L2→L1 accent/voice conversion

Twenty participants rated the accent and identity of the AC21 and VC21 models on a set of 12 sentences, randomly selected from the 50 sentences in the test set. Both models were perceived to be more foreign-accented than the original L1 utterances. On average, VC21 was rated as more foreign-accented than L1 (mean 97%; std. 8%), and AC21 was rated as more foreign-accented than L1 as well (mean 97%; std. 8%). More importantly, when compared against each other AC21 was rated as more foreign-accented than VC21 (mean 64%, std. 15%) which was statistically significant ($t = 4.3$, $p < 0.001$, single-tail).

Results from the ABX test show that the voice identity of AC21 was found to be more similar to L1 than to L2 (mean 67%; std. 28%), which is statistically significant ($t = 2.67$, $p = 0.008$, single-tail) compared to chance levels (50%).

In summary, these results show that the proposed AC method is also more effective than the baseline VC method in imparting a non-native accent to a native speaker, while it also preserves the identity of the L1 speaker. Results are summarized in Table 3.

Table 3: Summary of perceptual results; percentage denotes preference for the first item in the pair (second item in ABX); SA: source accent; ID: speaker identity.

	L1→L2	Pref.		L2→L1	Pref.
SA	VC12-L2	90%		VC21-L1	97%
	AC12-L2	89%		AC21-L1	97%
	AC12-VC12	60%		AC21-VC21	64%
ID	AC12-L2-L1	77%		AC21-L1-L2	67%

4.3.3 Correlation with differences in the L1 and L2 phonetic inventories

As a final step, we analyzed whether the effectiveness of the AC model could be explained from differences in the phonetic inventory of the two languages (Goldstein, 2001; Helman, 2004; You *et al.*, 2005). In particular, the English language includes a number of consonants that do not exist in Spanish, most significantly the fricatives /v/, /z/, /θ/, /ʃ/, /ʒ/ and /ð/, the affricate /dʒ/, the pseudo-fricative /h/, and the liquid /ɹ/. Spanish also does not have lax vowels, the schwa as well as r-colored vowels.

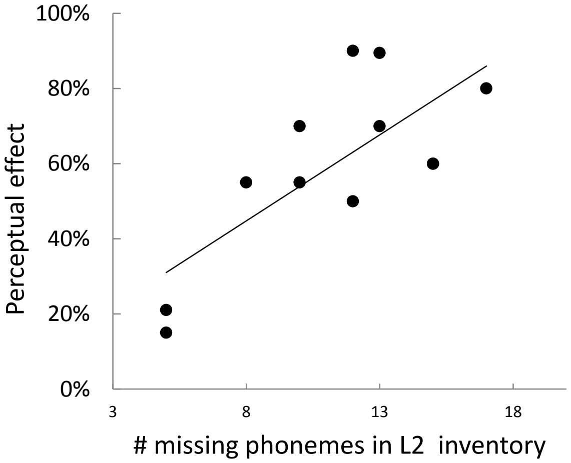


Figure 12: The number of missing phonemes in L2 inventory ($N_{p \notin L2}$) and the proportion of listeners who found the AC12 synthesis less foreign accented than the VC12 synthesis for each test sentence are highly correlated ($\rho = 0.74$).

Thus, for each sentence in the listening tests we computed the number of phonemes that did not exist in Spanish ($N_{p \notin L2}$), our rationale being that the larger this number the more difficult it would be for the L2 speaker to pronounce the sentence. Then, we computed the correlation coefficient between $N_{p \notin L2}$ and the proportion of listeners who found the AC12 synthesis less accented than the VC12 synthesis. Results reveal a very strong correlation ($\rho = 0.74$) between both measures (see Figure 12), which indicates that the benefits of the AC method are more significant for sentences that are harder to produce by the L2 speaker.

We also computed the correlation between $N_{p \notin L2}$ and the proportion of listeners who found AC12 less accented than L2; in this case, the correlation was $\rho = 0.63$, which adds further support to the previous conclusion. In contrast, the performance of the baseline voice conversion method (VC12) appears to be unrelated ($\rho = 0.07$) to the difficulty of the test sentence.

4.4 Discussion

This section has presented a speech modification method that can be used to transform L2 utterances to sound more native-accented. The method is based on conventional GMM techniques for voice conversion, but uses a different strategy to match frames from the source (L1) and target (L2) speakers. Namely, we apply vocal tract length normalization and then perform a bidirectional match between frames of the two speakers using Mel Cepstral Distortion as a measure of similarity; the resulting lookup table of source-target vectors is then used to train a GMM.

To test the effectiveness of our method, we compared it against a baseline voice-conversion model trained on DTW-aligned pairs of source-target utterances. Listening tests show that our accent conversion method can transfer the accent of the source speaker more effectively than voice conversion, regardless of the direction in which the transformations are applied, i.e., making L2 utterances less foreign-accented as well as making L1 utterances more foreign-accented.

Our results also show that the accent conversion method is most beneficial when used on utterances that are difficult to produce by L2 speakers, as measured by the number of phones in the utterance that do not exist in the L2 phonetic inventory. Further insights may be obtained by analyzing phonotactic differences between the two languages. A classic example in Spanish is the lack of word-initial clusters that begin with /s/; in these cases, Spanish speakers tend to produce such words (e.g., star, scar, small, Spain) with an initial /e/. One may also consider whether the particular error has high or low functional load (its importance in making distinctions in the language); as an example, contrast between initial /p/-/b/ has a high relative functional load, whereas final /t/-/d/ has a lower functional load (Jesse, 2012).

Further improvements in the accent conversion model may also be obtained by imposing constraints on the pairing of acoustic vectors. As an example, one may eliminate source-target pairs that have high Mel Cepstral Distortion. Performance may also be improved by considering additional information when matching source-target pairs, such as dynamic features (delta and delta-delta), features from the STRAIGHT aperiodicity spectrum, or linguistic features predicted from speech acoustics such as sound classes (e.g., place and manner of articulation).

In this section, we presented an acoustic-based strategy for foreign accent conversion. However, this dissertation focuses of developing foreign accent conversion techniques that exploits the voice-independent representation of linguistic gestures captured in articulatory data. In the next section, we will describe how the articulatory data provides a straightforward mechanism to transfer linguistic gestures (including accents and speaking styles) across speakers, and facilitates foreign accent conversion.

5. STATISTICAL PARAMETRIC ARTICULATORY FOREIGN ACCENT CONVERSION[9]

In this section, we present an articulatory-based strategy for foreign accent conversion. Unlike the acoustic-based approach, the articulatory-based approach has a strong theoretical basis. According to the modulation theory of speech (Traunmüller, 1994) speech is viewed as the result of process in which a carrier, characterized by the static properties of the speaker's voice, has been modulated by phono-articulatory gestures to give linguistic color. Based on this view, a speech synthesizer driven by articulatory gesture can be treated as the voice-quality carrier while the input articulatory gestures provide the modulating signal. Given the input articulatory gestures from a reference native speaker (L1) and the articulatory synthesizer built for a non-native speaker (L2), we can generate native-like utterance in the voice of the non-native speaker.

The method consists of building an articulatory synthesizer of the L2 speaker, then driving it with articulatory gestures[10] from an L1 speaker. As shown in Figure 13, the approach requires (i) a flexible articulatory synthesizer that can capture subtle accent-related changes in articulators, and (ii) an articulatory normalization method that

[9] The description of the method and the experimental results are reprinted with permission from "Reduction of non-native accents through statistical parametric articulatory synthesis," by Aryal and Gutierrez-Osuna, 2015. *J. Acoust. Soc. Am.*, 137, pp. 433-446. ©2015 Acoustical Society of America.

[10] We used the term articulatory gestures in a broader sense to represent the dynamics of vocal tract configurations. Not to be confused with 'gestures' and 'gestural scores' in the gestural framework of articulatory phonetics developed at Haskins Laboratories (Browman and Goldstein, 1990).

can account for physiological differences between the two speakers. This approach builds on a prior work on data-driven articulatory synthesis (Felps *et al.*, 2012), which illustrated the limitations of unit-selection techniques when used with small articulatory corpora[11]. For this reason, the method proposed here uses the Gaussian mixture model of Toda *et al.* (2008) to generate a forward mapping from L2 articulators to L2 acoustics. Compared to unit selection, this statistical parametric articulatory synthesizer does not require a large articulatory corpus and provides a continuous mapping from articulators to acoustics, so it can interpolate phonemes that do not exist in L2 inventory.

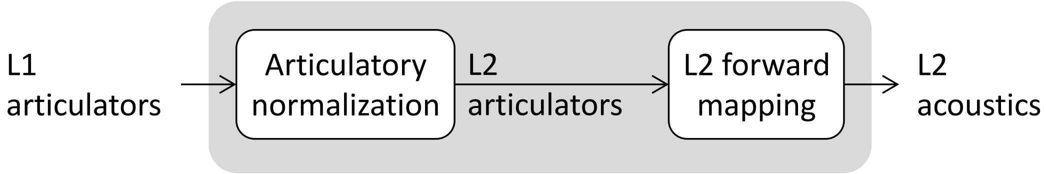


Figure 13: Articulatory accent conversion is a two-step process consisting of L1-L2 articulatory normalization and L2 forward mapping.

Given the differences in vocal tract physiology between the two speakers and in articulatory measurement procedures (e.g., pellet placement in electromagnetic articulography, or EMA), driving the resulting model with L1 articulators is unlikely to produce intelligible speech. To address this issue, Felps *et al.* (2012) mapped L1 and L2 articulators (EMA positions) into the 6-point Maeda parameter approximations of Al Bawab et al. (2008). While this parameterization can reduce individual speaker

[11] Previous work on text-to-speech unit-selection synthesis shows that at least two hours of active speech are needed to synthesize intelligible speech, a number that is rarely (if ever) achieved with articulatory corpora.

differences, it also reduces synthesis quality because it removes important information available in the raw EMA positions. For this reason, we achieve articulatory normalization by transforming EMA articulators between the two speakers by means of a pellet-dependent Procrustes transformation derived from articulatory landmarks of the two speakers, as proposed by Geng and Mooshammer (2009).

5.1 Method description

Our proposed articulatory method for accent conversion follows the generic outline shown in Figure 13. The method takes an acoustic-articulatory trajectory from an L1 test utterance and transforms it to match the voice quality of the L2 speaker. In a first step, the method normalizes the L1 articulatory trajectory (EMA pellet coordinates) to the L2 articulatory space. Then, it uses the normalized L1 trajectories as an input to a GMM-based articulatory synthesizer trained on an L2 acoustic-articulatory corpus. The result is an utterance that has the articulatory gestures and prosody of the L1 speaker but the voice quality of the L2 speaker. Both procedures are described in detail in the following sections.

5.1.1 Cross-speaker articulatory mapping

The articulatory mapping transforms a vector $\boldsymbol{x_{L1}}$ of EMA articulatory coordinates for the L1 speaker into the equivalent articulatory positions $\boldsymbol{\hat{x}_{L2}} = f_{12}(\boldsymbol{x_{L1}})$, where $f_{12}(\cdot)$ denotes a set of Procrustes transforms, one for each fleshpoint. Namely, given an L1 fleshpoint with antero-posterior and supero-inferior coordinates $(x_{L1,a}, x_{L1,s})$, the function estimates the L2 fleshpoint coordinates $(\hat{x}_{L2,a}, \hat{x}_{L2,s})$ as:

$$[\hat{x}_{L2,a}, \hat{x}_{L2,s}] = [c_a, c_s] + \rho\,[x_{L1,a}, x_{L1,s}]\,\boldsymbol{A} \tag{18}$$

where $[c_a, c_s]$ is the translation vector, ρ is the scaling factor and $\boldsymbol{A}$ is a 2×2 matrix representing the rotation and reflection. We estimate the Procrustes parameters$\{c_{\boldsymbol{a}}, c_s, \rho, \boldsymbol{A}\}$ by solving the minimization problem:

$$\min_{\{c_a, c_s, \rho, \boldsymbol{A}\}} \sum_{all\ landmarks} \left\| [x_{L2,a}, x_{L2,s}] - \left([c_a, c_s] + \rho\,[x_{L1,a}, x_{L1,s}]\,\boldsymbol{A}\right) \right\| \tag{19}$$

where $[x_{L2,a}, x_{L2,s}]$ and $[x_{L1,a}, x_{L1,s}]$ are the coordinates of corresponding landmarks in the L2 and L1 speaker respectively. These parameters are learned for each pellet in the articulatory corpus.

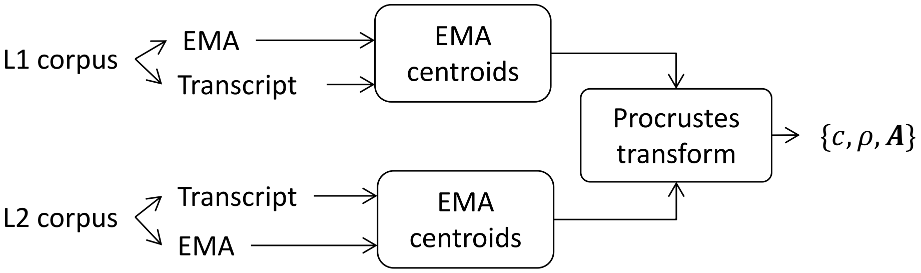


Figure 14: Overview of the cross-speaker articulatory normalization procedure. A separate set of parameters is obtained for each EMA pellet.

Following Geng and Mooshammer (2009), we select a set of articulatory landmarks from the phonetically-transcribed corpus. Namely, for each phone in the L1 inventory and for each speaker, we calculate the centroid of the EMA articulatory coordinates as the average across all frames that belong to the phone (according to the phonetic transcription). These pairs of phone centroids (one from the L1 speaker, one from the L2 speaker) are then used as the corresponding landmarks in equation (19). The overall approach is summarized in Figure 14.

5.1.2 Forward mapping

To generate acoustic observations from articulatory positions, we use a GMM-based forward mapping (Toda *et al.*, 2008) that incorporates global variance (GV) of the acoustic features (Toda *et al.*, 2007). The forward mapping estimates the temporal sequence of static acoustic parameters, ($MFCC_{1-24}$), from the trajectory of articulatory features $\boldsymbol{x}$. For each frame at time t, the articulatory feature vector $\boldsymbol{x_t}$ consists of 15 parameters: the anteroposterior and superoinferior coordinate of six EMA pellets, pitch ($\log f_0$), loudness ($MFCC_0$) and *nasality*. Since the velum position is not available in our EMA corpus, we used the text transcription of the utterances to generate a binary feature that represented *nasality*. In the absence of a transcription, the nasality feature may be derived from acoustic features –see (Pruthi and Espy-Wilson, 2004), as in the case for fundamental frequency and loudness.

For completeness, we include a detailed description of the forward mapping in (Toda *et al.*, 2007; Toda *et al.*, 2008). In a first step, we model the joint distribution of articulatory-acoustic features $\boldsymbol{Z_t} = [\boldsymbol{x}_t, \boldsymbol{Y}_t]$, where $\boldsymbol{x_t}$ is the articulatory feature vector at time t, and $\boldsymbol{Y}_t = [\boldsymbol{y}_t, \Delta\boldsymbol{y}_t]$ is an acoustic feature vector containing both static and delta MFCCs. Using a Gaussian mixture, the joint distribution becomes:

$$p\left(\boldsymbol{Z_t}\middle|\boldsymbol{\lambda}^{(z)}\right) = \sum_{m=1}^{M} \alpha_m \, \mathcal{N}(\boldsymbol{Z_t}; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}) \tag{20}$$

where α_m is the scalar weight of the m^{th} mixture component and $\mathcal{N}(; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)})$ is the Gaussian distribution with mean $\boldsymbol{\mu}_m^{(z)}$and covariance matrix $\boldsymbol{\Sigma}_m^{(z)}$. We use symbol $\boldsymbol{\lambda}^{(z)} = \{\alpha_m, \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}\}$ to denote the full parameter set for the GMM. The mean vector $\boldsymbol{\mu}_m^{(Z)}$ and covariance matrix $\boldsymbol{\Sigma}_m^{(Z)}$ denote the joint statistics of articulatory and acoustic features for the m^{th} mixture:

$$\boldsymbol{\mu}_m^{(Z)} = \left[\boldsymbol{\mu}_m^{(x)} \ \boldsymbol{\mu}_m^{(Y)}\right], \boldsymbol{\Sigma}_m^{(Z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xY)} \\ \boldsymbol{\Sigma}_m^{(Yx)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}. \tag{21}$$

In a second step, we model the global variance (GV) of predicted acoustics to account for over-smoothing effects of the GMM. Consider the within-sentence variance of the d^{th} acoustic feature $y_t(d)$, given by $v(d) = E[(y_t(d) - E[y_t(d)])^2]$. The GV of these features in an utterance $\boldsymbol{y}\,(= [\boldsymbol{y}_1, \boldsymbol{y}_2 \dots \boldsymbol{y}_T])$ is then given by a vector $\boldsymbol{v}(\boldsymbol{y}) = [\,v(1), v(2) \dots v(D)]$, where D is the dimension of acoustic vector $\boldsymbol{y}_t$. We model the distribution of GVs for all the utterances in the training set, $P\left(\boldsymbol{v}(\boldsymbol{y})\middle|\boldsymbol{\lambda}^{(v)}\right)$, with a single Gaussian distribution:

$$p\left(\boldsymbol{v}(\boldsymbol{y})\middle|\boldsymbol{\lambda}^{(v)}\right) = \mathcal{N}\left(\boldsymbol{v}(\boldsymbol{y}); \boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(vv)}\right) \tag{22}$$

where model parameters $\boldsymbol{\lambda}^{(v)} = \left\{\boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(vv)}\right\}$ denote the vector of average global variance $\boldsymbol{\mu}^{(v)}$ and the corresponding covariance matrix $\boldsymbol{\Sigma}^{(vv)}$, learned from the distribution of $\boldsymbol{v}(\boldsymbol{y})$ in the training set.

At synthesis time, given the trained models $[\boldsymbol{\lambda}^{(z)}, \boldsymbol{\lambda}^{(v)}]$ and a test sequence of articulatory vectors $\boldsymbol{x} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3} \dots \boldsymbol{x_T}]$**,** we obtain the maximum-likelihood acoustic (static only) trajectory $\widehat{\boldsymbol{y}}$:

$$\widehat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}} \ P\big(\boldsymbol{Y}\big|\boldsymbol{x}, \boldsymbol{\lambda}^{(z)}\big)^{\omega} . P\big(\boldsymbol{v}(\boldsymbol{y})\big|\boldsymbol{\lambda}^{(v)}\big) \tag{23}$$

where $\boldsymbol{Y} = [\boldsymbol{y_1}, \Delta\boldsymbol{y_1}, \boldsymbol{y_2}, \Delta\boldsymbol{y_2}, \dots \boldsymbol{y_T}, \Delta\boldsymbol{y_T}\ ]$ is the time sequence of acoustic vectors (both static and dynamic) and $\boldsymbol{v}(\boldsymbol{y})$ is the variance of static acoustic feature vectors. The power term $\omega\ (= 1/2T)$ in equation (23) provides a balance between the two likelihoods. We solve for $\widehat{\boldsymbol{y}}$ in equation (23) via Expectation-Maximization; for details refer to section 4.1.3.

5.1.3 System diagram

Figure 15 shows a more detailed view of the conversion process. The system takes audio and articulatory recordings from a reference L1utterance as the input. From the audio signal, STRAIGHT extracts pitch, aperiodicity and energy. Given the trained pitch modification module as described in section 4.1.4, the L1 pitch trajectory is shifted and scaled to match the pitch range of the L2 speaker.

Given the trained cross-speaker articulatory mappings (described in section 5.1.1), the L1 articulatory trajectories (EMA pellet positions) are transformed to the equivalent L2 articulatory trajectories. The frame energy ($MFCC_0$) and the modified pitch (p_{L2}) are combined with the normalized L1 EMA positions and the binary nasality feature to form an input articulatory feature vector for the L2 forward mapping (described in section 0), which estimates the L2 spectral coefficients ($MFCC_{1-24}$).

We then reconstruct the STRAIGHT *spectral envelope* from the estimated L2 spectral coefficients (*$MFCC_{1\text{-}24}$*) and the L1 energy (*$MFCC_0$*). In a final step, we use the STRAIGHT synthesis engine to generate the waveform using the estimated *spectral envelope* $\hat{\boldsymbol{s}}$, the L1 *aperiodicity* and the modified *pitch*.

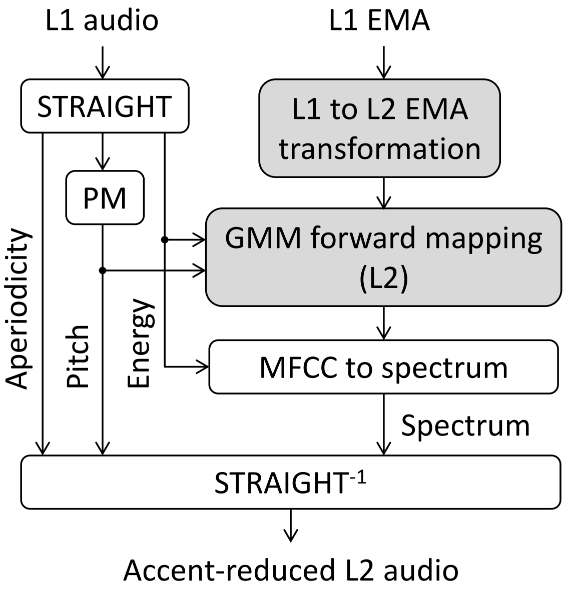


Figure 15: Block diagram of accent conversion method (PM: pitch modification).

5.2 Experimental corpus

We performed a series of perceptual listening experiments to evaluate the proposed method in terms of its ability to improve intelligibility, reduce non-native accentedness, and preserve voice individuality. For this purpose, we used a corpus of audio and articulatory recordings from a native speaker of American English, and a non-native speaker whose first language was Spanish (Felps *et al.*, 2012; Aryal and

Gutierrez-Osuna, 2013) collected at the University of Edinburgh by means of Electromagnetic Articulography (EMA; Carstens AG500). Both speakers recorded the same 344 sentences chosen from the Glasgow Herald corpus. The non-native speaker recorded an additional 305 sentences from the same corpus. Out of the 344 common sentences, we randomly selected 50 sentences (220 seconds of active speech in total; 4.40 seconds/sentence on average) for testing, and used the remaining 294 sentences (1,290 seconds total; 4.39 seconds/sentence) to train the forward mapping and the articulatory mapping. Six standard EMA pellets positions were recorded: upper lip, lower lip, lower jaw, tongue tip, tongue body, and tongue dorsum. Four additional pellets (placed behind the ears, the upper nasion and the upper jaw) were used to cancel head motion and provide a frame of reference. EMA pellet positions were recorded at 200Hz. From each acoustic recording, we also extracted pitch, aperiodicity and spectral envelope using STRAIGHT (Kawahara, 1997). MFCCs were then estimated from the STRAIGHT spectrum and resampled to match the EMA recordings. The result was a database of articulatory-acoustic feature vectors containing pitch, $MFCC_{0-24}$ and six EMA positions per frame.

5.2.1 Experimental conditions

We considered five different experimental conditions for the listening tests: the proposed accent conversion method (AC), articulatory synthesis of L2 utterances ($L2_{EMA}$), articulatory synthesis of L1 utterances ($L1_{EMA}$), MFCC compression of L2 speech ($L2_{MFCC}$), and normalization of L1 utterances to match the vocal tract length and pitch range of L2 ($L1_{GUISE}$). The conditions are summarized in Table 4.

Table 4: Five experimental conditions for listening test.

Experimental condition	Aperiodicity and energy	Pitch	Articulators	Spectrum
AC	L1	L1 scaled to L2	L1 mapped to L2	L2 forward mapping
$L2_{EMA}$	L2	L2	L2	L2 forward mapping
$L1_{EMA}$	L1	L1	L1	L1 forward mapping
$L2_{MFCC}$	L2	L2	N/A	L2 MFCC
$L1_{GUISE}$	L1	L1 scaled to L2	N/A	L1 warped to L2

The first experimental condition (AC) was the proposed accent conversion method, illustrated in Figure 13. Namely, we built an L2 forward mapping by training a GMM with 128 mixtures on L2 articulatory-acoustic frames, and the Procrustes articulatory registration model by training on the articulatory landmarks of equation (19); only non-silent frames in the 294 training sentences were used for this purpose. Once the cross-speaker articulatory mapping and L2 forward mapping had been trained, we performed accent-conversion for each of the L1 utterances not used for training, following the procedure outlined in Figure 15.

The second experimental condition ($L2_{EMA}$) consisted of articulatory synthesis of L2 utterances, obtained by driving the L2 forward model with L2 articulators. This condition was used as the baseline for non-native production of the utterances since it had similar acoustic quality as AC. Because articulatory synthesis results in a loss of acoustic quality, and considering that acoustic quality interacts with accent perception

(Felps *et al.*, 2010), comparing AC against the L2 original utterances would have been problematic.

The third experimental condition ($L1_{EMA}$) consisted of articulatory synthesis of L1 utterances, obtained by driving an L1 forward model with L1 articulators. This condition served as the baseline for native production of the utterances, accounting for the loss of quality due to articulatory synthesis. This condition may also be taken as an upper bound of what accent conversion may be able to achieve in terms of intelligibility and accentedness.

The fourth experimental condition ($L2_{MFCC}$) consisted of re-synthesizing the original L2 utterances following compression into MFCCs. Utterances in this condition underwent a four-step process: (1) STRAIGHT analysis, (2) compression of STRAIGHT smooth spectra into MFCCs, (3) reconstruction of STRAIGHT smooth spectra from MFCCs, and (4) STRAIGHT synthesis; refer to section 4.1.1 for more detail on steps (2) and (3). This modification enabled a fair comparison against AC utterances by factoring out losses in acoustic quality caused by the MFCC compression step in Figure 15.

The fifth experimental condition ($L1_{GUISE}$) consisted of modifying L1 utterances to match the pitch range and vocal tract length of the L2 speaker. This condition allowed us to test whether a simple guise could achieve similar accent-conversion performance as the proposed AC method: as shown in a number of studies (Lavner et al., 2000, and references therein) , pitch range and formant frequencies are good indicators of voice identity. Utterances in the $L1_{GUISE}$ condition were synthesized as follows. First, the L1 pitch trajectory was rescaled and shifted to match the pitch range of L2 speaker(17).

Then, we performed vocal tract length normalization by warping the L1 STRAIGHT spectrum to match the global statistics of the L2 speaker. Following Sundermann *et al.* (2003), we used a piecewise linear warping function governed by the average formant pairs of the two speakers, estimated over the training corpus; formants were extracted from the roots of the LPC coefficients of non-silent frames. For similar reasons as those described above, $L1_{GUISE}$ utterances also underwent the same MFCC compression procedure of $L2_{MFCC}$ utterances.

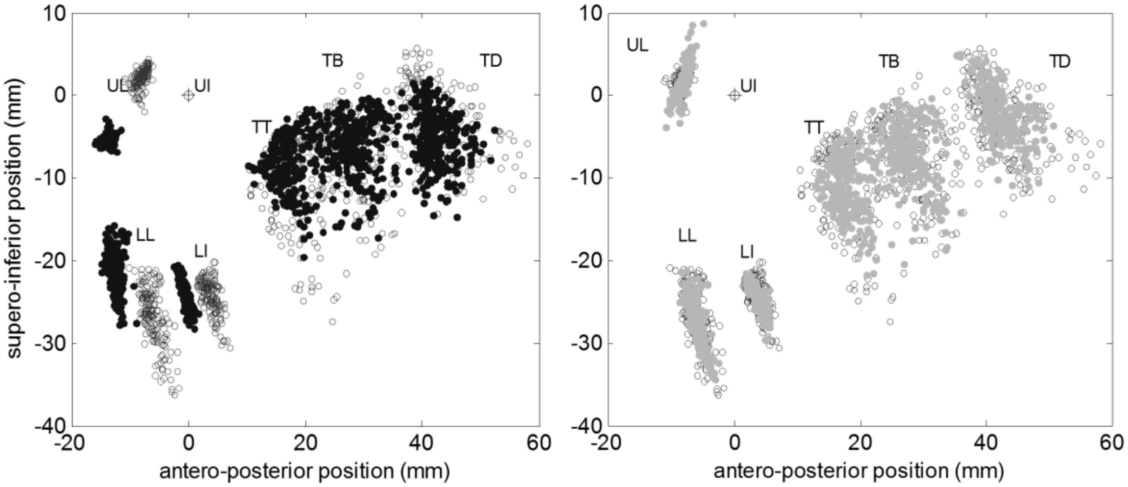
5.2.2 Participant recruitment

We evaluated the proposed method (AC) by comparing against the other four experimental conditions ($L2_{EMA}$, $L1_{EMA}$, $L2_{MFCC}$, $L1_{GUISE}$) in terms of intelligibility, accentedness and speaker individuality through a series of perceptual listening tests. Participants for the perceptual studies were recruited through Mechanical Turk, Amazon's online crowdsourcing tool. In order to qualify for the studies, participants were required to reside in the United States and pass a screening test that consisted of identifying various American English accents: Northeast (i.e. Boston, New York), Southern (i.e. Georgia, Texas, Louisiana), and General American (i.e. Indiana, Iowa). Participants who did not pass this qualification task were not allowed to participate in the studies. In addition, participants were asked to list their native language/dialect and any other fluent languages that they spoke. If a subject was not a monolingual speaker of American English then their responses were excluded from the results. In the quality and accent evaluation tests, participants were asked to transcribe the utterances to ensure they paid attention to the recordings. Participants with incomplete responses were excluded from the study.

5.3 Results

5.3.1 Accuracy of articulatory normalization

In a first experiment, we analyzed the effect of the Procrustes transforms on the distribution of articulatory configurations. First, we compared the spatial distribution of the six EMA pellets for the L1 and L2 speakers before and after articulatory normalization. Figure 16a shows the distribution before articulatory normalization; differences between the two speakers are quite significant, not only in terms of the average position of each pellet but also in terms of its spatial distribution (e.g., variance). These discrepancies can be attributed largely to differences in vocal tract geometry between the two speakers, though inconsistencies in pellet placement during the EMA recordings also play a role. Regardless of the source of these discrepancies, the results in Figure 16b shows that the articulatory normalization step achieves a high degree of consistency in the spatial distribution of pellets between the two speakers.



(a) (b)

Figure 16: (a) Distribution of six EMA pellet positions from the L1 speaker (solid markers) and L2 speaker (hollow markers) from a parallel corpus. Large differences can be seen in the span of the measured positions of articulators (UL: upper lip; LL: lower lip; LI: lower incisor; TT: tongue tip; TB: tongue blade; and TD: tongue dorsum). The upper incisor (UI) was used as a reference point. (b) Distribution of EMA pellet positions for the L1 speaker (solid markers) and L2 speaker (hollow markers) following articulatory normalization.

Next, we compared articulatory trajectories for the L1 speaker, the L2 speaker, and the L1 after articulatory normalization. Figure 17 shows the trajectory of tongue tip for the word ‘that’ in a test utterance. As a result of the normalization step, the L1 articulatory trajectory becomes closer to the L2 trajectory but also preserves the dynamics of the L1 production; this makes it easier to spot articulatory errors in the L2 utterance. Namely, the figure shows a noticeable difference between the L2 trajectory and the L1-normalized trajectory in antero-posterior position towards the end of the word. This discrepancy can be traced back to a typical phonetic substitution of alveolar stop /t/ with the dental one /t̪/ in L2 speakers whose mother tongue is Spanish, which results from moving the tongue tip forward to make a constriction at the teeth instead of the alveolar ridge. Such display of normalized trajectories may also be used as supplementary feedback mechanism to the learner in computer-assisted pronunciation training.

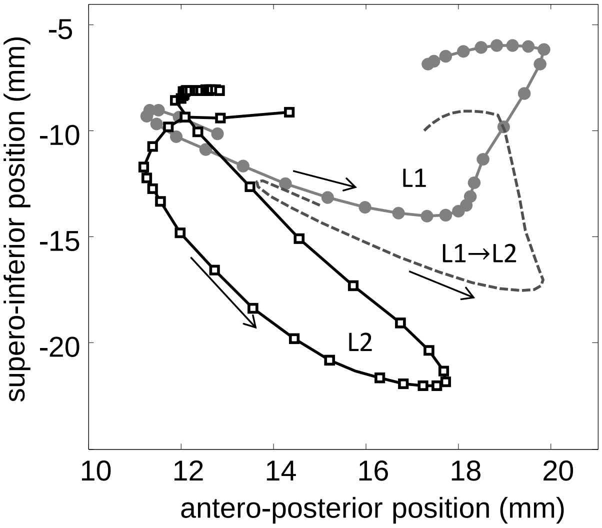


Figure 17: Trajectory of the tongue-tip pellet in L1 and L2 utterances of the word ‘that’. The L1 trajectory normalized to the L2 articulatory space is also shown. Arrows indicate the direction of trajectories.

Finally, we analyzed the effect of articulatory normalization on the distribution of articulatory configurations at the phonetic level; the middle frame of vowel segments was used for this purpose. Figure 18a shows the centroid and half-sigma contour (i.e., half standard deviation) of the tongue tip pellet position, a critical articulator for the frontal vowels (/ɪ/, /i/, /ɛ/, /e/ and /æ/), for the two speakers (L1 and L2). As shown in Figure 18a, the half-sigma contours for corresponding vowels in the two speakers have no overlap, with the exception of /ɪ/ and /ɛ/. Notice also the larger spread in articulatory configurations for the L2 speaker compared to the L1 speaker, a result that is consistent with prior studies showing larger acoustic variability and phonemic overlap in non-native speech productions (Wade *et al.*, 2007). Figure 18b shows the articulatory configurations following the articulatory normalization step; vowel centroids for the normalized L1 speaker are within the half-sigma contour of the corresponding vowel for the L2 speaker.

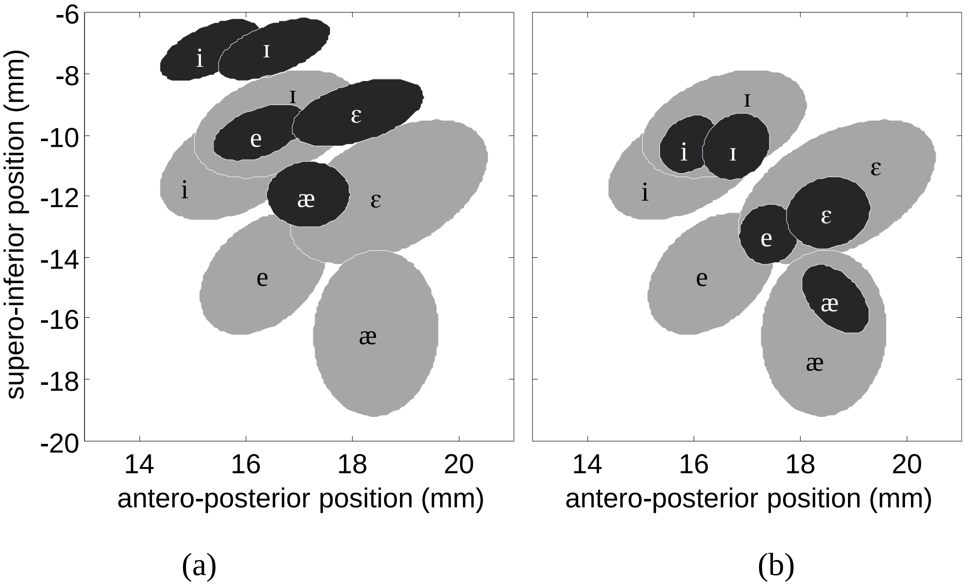
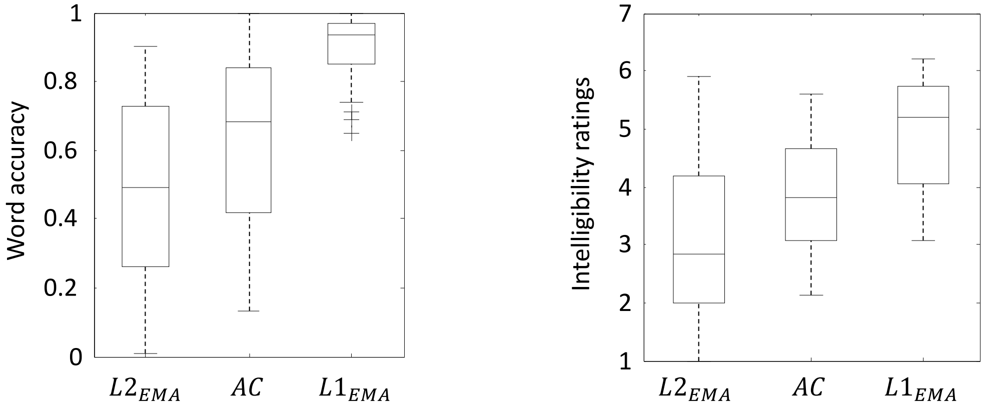


Figure 18: (a) Distribution of tongue tip position in frontal vowels for the L1 speaker (dark ellipses) and L2 speaker (light) speaker; ellipses represent the half-sigma contour of the distribution for each vowel. (b) Distribution of tongue tip position in frontal vowels for the L1 speaker after articulatory mapping (dark) and the L2 speaker (light).

5.3.2 Assessment of intelligibility

In a first listening test we assessed the intelligibility of AC as compared to $L1_{EMA}$ and $L2_{EMA}$ utterances. Three independent groups of native speakers of American English

(N=15 each) transcribed the 46 test utterances[12] for the three experimental conditions (AC, $L1_{EMA}$, $L2_{EMA}$). From each transcription, we calculated word accuracy (W_{acc}) as the ratio of the number of correctly identified words to the total number of words in the utterance. Participants also rated the (subjective) intelligibility of the utterances (S_{intel}) using a 7-point Likert scale (1: not intelligible at all, 3: somewhat intelligible, 5: quite a bit intelligible, and 7: extremely intelligible).



(a) (b)

Figure 19: Box plot of (a) word accuracy and (b) subjective intelligibility ratings for $L1_{EMA}$, $L2_{EMA}$ and AC utterances.

Figure 19 shows the word accuracy and intelligibility ratings for the three experimental conditions. Accent conversions (AC: $W_{acc} = 0.64$, $S_{intel} = 3.83$) were rated as being significantly more intelligible ($p < 0.01; t - test$) than L2 articulatory synthesis ($\mathrm{L2_{EMA}}$: $W_{acc} = 0.50$, $S_{intel} = 3.30$), a result that supports the feasibility of

[12] Four of 50 test sentences for the L2 speaker had missing EMA data and were removed from the analysis.

our accent-conversion approach, though not as intelligible ($p < 0.01; t - test$) as the upper bound of L1 articulatory synthesis ($L1_{EMA}$: $W_{acc} = 0.90$, $S_{intel} = 4.96$). In all three conditions, the two intelligibility measures (W_{acc}, S_{intel}) were found to be significantly correlated ($\rho > 0.89, N = 46$); for this reason, in what follows we will focus on W_{acc} as it is the more objective of the two measures.

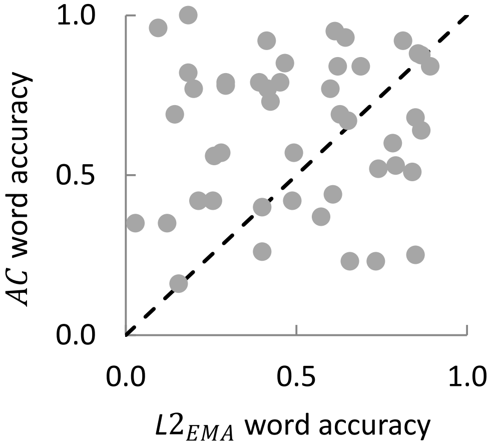


Figure 20: Word accuracy for AC and $L2_{EMA}$ for the 46 test sentences. The diagonal dashed line represents The sentences for which $W_{acc}(AC) > W_{acc}(L2_{EMA})$ are above the dashed line and the vice versa.

The scatter plot in Figure 20 shows the AC and $L2_{EMA}$ word accuracies for the 46 test sentences. In 70% of the cases (32 sentences; those above the main diagonal in the figure) accent conversion improved word accuracy compared to that obtained on $L2_{EMA}$ utterances, further supporting our approach. Notice, however, the lack of correlation between the two conditions, an unexpected result since one would expect that the initial word accuracy (i.e., on $L2_{EMA}$ utterances) would have a strong influence on word accuracy following accent conversion. As will be discussed next, this result suggests the presence of two independent factors affecting intelligibility in the two conditions.

The results in Figure 19a also show a large variance in word accuracy for L2 articulatory synthesis ($L2_{EMA}$) compared to L1 articulatory synthesis ($L1_{EMA}$). In our analysis of the acoustic-based accent conversion in section 4.3.3, we found that accent conversions are most beneficial when used on utterances that are difficult to produce by L2 speakers based on differences between the L1 and L2 phonetic inventories. Accordingly, we examined whether the variance in word accuracy for $L2_{EMA}$ could be explained by the phonetic complexity of each sentence, measured as the number of L1 phones in the sentence that do not exist in the L2 inventory. Differences in phonetic inventories are a known reason behind non-native accents; see e.g. (You *et al.*, 2005). In our case, the English language includes a number of consonants that do not exist in Spanish (our L2 speaker's mother tongue), most significantly the fricatives /v/, /z/, /θ/, /ʃ/, /ʒ/ and /ð/, the affricate /ĵ/, the pseudo-fricative /h/, and the liquid /ɹ/. Spanish also does not have lax vowels, the schwa as well as r-colored vowels. Thus, for each test sentence we computed the number of phones that did not exist in Spanish ($N_{p\notin L2}$) and compared it against the $L2_{EMA}$ word accuracy. Both variables ($N_{p\notin L2}$, $W_{acc}(L2_{EMA})$) are significantly correlated ($\rho = 0.44, N = 46, p < 0.01$) , suggesting that variance in intelligibility for $L2_{EMA}$ utterances can be explained by differences in the L1 and L2 phonetic inventory. We found, however, no significant correlation ($\rho = -0.2; p = 0.09$) between $N_{p\notin L2}$ and word accuracy for AC utterances, which suggests that the accent conversion process is able to cancel out the main source of (poor) intelligibility: phonetic complexity from the perspective of the L2 learner.

What then, if not sentence complexity, drives the intelligibility of AC utterances? Since both conditions $(AC, L2_{EMA})$ use the same articulatory synthesizer, we hypothesized that interpolation issues would be at fault. To test this hypothesis, for each frame in an AC utterance we computed the Mahalanobis distance between the L1 registered articulators and the centroid of the corresponding L2 phone, then averaged the distance over all non-silent frames in the utterance. The larger this measure, the larger the excursion of the registered L1 articulatory trajectory from the L2 articulatory space. We found, however, no significant correlation $(\rho = -0.21; p = 0.08)$ between this measure and word accuracy on AC utterances, which suggests that the total amount of interpolation present in an AC utterance does not explain its lack of intelligibility.

Table 5: Correlation between word accuracy and the proportion of phones in a sentence containing a particular articulatory-phonetic feature.

	Articulatory features	$\boldsymbol{AC}$	$\boldsymbol{L2_{EMA}}$	$\boldsymbol{L1_{EMA}}$
Manner	Stops	**-0.43**	0.22	-0.21
	Fricatives	-0.01	0.04	-0.02
	Affricates	0.05	-0.17	0.05
	Nasals	0.31	-0.10	0.17
	Liquids	-0.19	-0.08	-0.22
	Glides	**0.40**	0.01	0.17
Place	Bilabials	-0.07	0.28	-0.07
	Labiodentals	0.14	-0.11	-0.03
	Lingual dental	-0.18	-0.03	-0.12
	Lingual alveolar	-0.04	-0.12	0.10
	Lingual palatal	0.02	-0.21	-0.04
	Lingual velar	0.01	0.25	-0.08
	Glottal	0.01	0.14	-0.09
Voicing	Voiced	0.01	-0.07	-0.18
	Unvoiced	-0.10	0.14	0.07

In a final analysis we then decided to test whether the phonetic content of the utterance would explain its intelligibility, our rationale being that the acoustic effect of interpolation errors is not uniform across phones. As an example, due to the presence of critical articulators, a small error in the tongue tip height can transform a stop into a fricative whereas the same amount of error in tongue tip height may not make much of a difference in a vowel. Accordingly, we calculated the correlation between word accuracy and the proportion of phones in an utterance with a specified phonetic-articulatory feature. Results are shown in Table 5 for six features of manner of articulation, seven features of place of articulation, and voicing. Correlation coefficients found to be significant ($p < 0.01$) are shown in bold. In the case of $L1_{EMA}$ and $L2_{EMA}$ utterances, we found no significant effect on intelligibility for any of the articulatory features, an indication that the GMM articulatory synthesizer was trained properly. In the case of AC utterances, however, we found a strong negative correlation between intelligibility and the proportion of stops in the sentence. Thus, it appears that small registration errors, to which stops are particularly sensitive, are largely responsible for the loss of intelligibility in accent-converted utterances[13].

5.3.3 Assessment of non-native accentedness

In a second listening experiment we sought to determine whether the proposed accent-conversion method could also reduce the perceived non-native accent of L2 utterances. For this purpose, participants were asked to listen to $L2_{EMA}$ and AC utterances of the same sentence and select the most native-like[14] among them. For this test, we focused on the subset of sentences for which AC and $L2_{EMA}$ utterances had higher intelligibility ($W_{acc} > 0.5$); i.e., those on the upper-right quadrant in Figure 20. In this way, we avoided asking participants to rate which of two unintelligible utterances was less foreign-accented (a questionable exercise) or whether an unintelligible

[13] The table also shows a strong positive correlation between intelligibility and glides, an unexpected result because it suggests that lowering the proportion of glides in an utterance reduces its intelligibility. A closer look at the phonetic composition of our 46 test utterances, however, shows that the proportion of glides is negatively correlated with the proportion of stops ($\rho = -0.32, p = 0.015$). This provides a more plausible explanation: as the proportion of glides decreases, so does the proportion of stops increase, in turn lowering the intelligibility of the utterance.

[14] Native relative to a monolingual speaker of general American English

utterance was more foreign-accented than an intelligible one (an exercise of predictable if not obvious results).

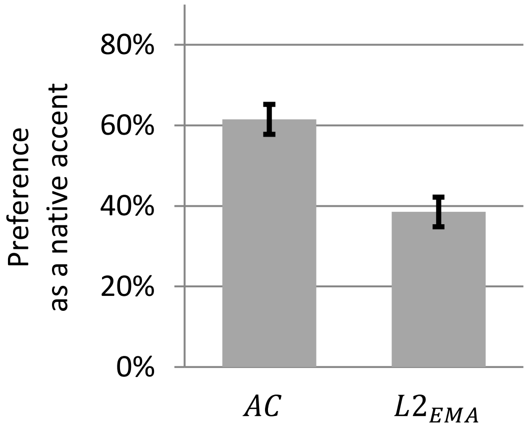


Figure 21: Subjective evaluation of non-native accentedness. Participants were asked to determine which utterance in a pair was more native-like.

Participants (N=15) listened to 30 pairs of utterances (15 $AC - L2_{EMA}$ pairs, and 15 $L2_{EMA} - AC$ pairs) presented in random order to account for order effects. Their preferences are summarized in Figure 21. AC utterances were rated as being more native than $L2_{EMA}$ utterances in 62% of the sentences (SE 4%), which is significantly higher than the 50% chance level $(t = 3.2, df = 14, p = 0.003, single\ tail)$. This result indicates the proposed accent-conversion method can be effective in reducing the perceived non-native accent of L2 utterances. To verify that these results were not accidental (e.g., caused by the lower acoustic quality of articulatory synthesis), we performed an additional listening test to compare accent ratings for native ($L1_{EMA}$) and non-native ($L2_{EMA}$) articulatory synthesis. In this test, a different group of listeners

(N=15) compared 30 pairs of utterances (15 $L1_{EMA} - L2_{EMA}$ pairs, and 15 $L2_{EMA} - L1_{EMA}$ pairs), and selected the most native-like utterance in a pair. As expected, $L1_{EMA}$ utterances were rated as more native than $L2_{EMA}$ in 96% of the cases, which indicates that articulatory syntheses do retain dialect/accent information.

Closer inspection of the listeners' responses to the accent perception comparisons showed an influence of presentation order within pairs. Namely, AC was rated as more native than $L2_{EMA}$ 53% of the times whenever AC appeared first, but the proportion increased to 70% if AC was the second utterance in the pair; this difference was statistically significant ($t = 4.0, df = 14, p < 0.001, single\ tail$). This bias is consistent with the 'pop-out' effect (Davis *et al.*, 2005), according to which a degraded utterance is perceived as being less degraded if presented after a clean version of the same utterance, i.e. when the lexical information is known. Extending this result to the perception of native accents, $L2_{EMA}$ may then be treated as the degraded utterances relative to the AC condition, which would explain why $L2_{EMA}$ utterances were rated as less accented if they were presented after AC.

5.3.4 Assessment of voice individuality

In a third and final listening experiment we tested the extent to which the accent conversion method was able to preserve the voice identity of the L2 speaker. For this purpose, we compared AC utterances against $L2_{MFCC}$ utterances (MFCC compressions of the original L2 recordings) and $L1_{GUISE}$ utterances (a simple guise of L1 utterances to match the vocal tract length and pitch range of the L2 speaker).

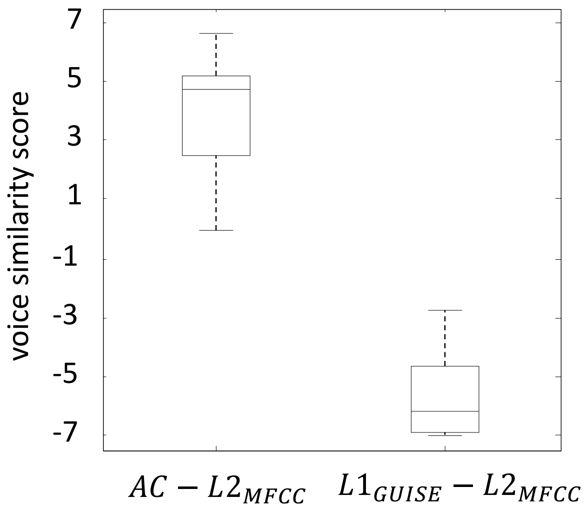


Figure 22: Average pairwise voice similarity scores. Scores range from -7 (different speaker with high confidence) to +7 (same speaker with high confidence).

Following Felps *et al.* (2009), we presented participants with a pair of linguistically different utterances from two of the three experimental conditions. Presentation order was randomized for conditions within each pair and for pairs of conditions. Participants (N=15) rated 40 pairs, 20 from each group ($AC - L2_{MFCC}$, $L1_{GUISE} - L2_{MFCC}$) randomly interleaved, and were asked to (1) determine if the utterances were from the same or a different speaker (forced choice), and (2) rate how confident they were in their assessment using a 7-point Likert scale. Once the ratings were obtained, participants' responses and confident levels were combined to form a *voice similarity score* (VSS) ranging from -7 (extremely confident they are different speakers) to $+7$ (extremely confident they are the same speaker).

Figure 22 shows the mean VSS between pairs of experimental conditions. Listeners were 'quite' confident that AC and $L2_{MFCC}$ utterances were from the same speaker ($VSS = 4.2, SE = 0.5$). This result suggests that the method is able to preserve the voice-identity of the L2 learner. Likewise, listeners were very confident ($VSS = -5.9, SE = 0.3$) that $L1_{GUISE}$ and $L2_{MFCC}$ utterances were from different speakers, which indicates that a simple guise of the L1 speaker cannot capture the voice quality of the L2 learner.

5.4 Discussion

This section has presented an accent-conversion method that transforms non-native utterances to match the articulatory gestures of a reference native speaker. Our approach consists of building a GMM-based articulatory synthesizer of a non-native learner, then driving it with measured articulatory gestures from a native speaker. Results from listening tests show that accent conversion provides statistically-significant increases in intelligibility as measured by objective scores (i.e. word recognition) and subjective ratings, and overall preference (70%) when compared to synthesis driven by L2 articulators. More importantly, unlike in the case of synthesis driven by L2 articulators, the intelligibility of accent conversions is not affected by the proportion of phones outside the phonetic inventory of the L2 speaker. This result suggests that the method can successfully remove one of the primary causes of non-native accents. Subsequent pairwise listening tests of native accentedness also show a preference towards accent conversions (62%) when compared to synthesis driven by L2 articulators. Finally, listening tests of speaker identity indicate that driving the L2 articulatory synthesizer with (registered) articulatory gestures from a different speaker does not change the perceived voice quality of the resulting synthesis. When combined with our results on intelligibility and accentedness, this finding suggests that our overall approach (L1→L2 articulatory normalization followed by L2 articulatory synthesis) is an effective strategy to decouple those aspects of an utterance that are due to the speaker physiology from those that are due to the language.

Further analysis indicates that the intelligibility of accent-converted utterances decreases with the proportion of stop consonants in the sentence. Given that stops require the formation of a complete constriction, small articulatory registration errors can have a significant effect on the acoustic output of the model; as an example, a small error in tongue-tip height may cause a lingua-alveolar stop to become fricative (e.g., from /t/ to /s/). A potential solution to this problem may be to incorporate knowledge of critical articulators by replacing the mapping in equation (18) with one that is context-dependent. To this end, Felps *et al.* (2010) have shown that the accuracy of articulatory-acoustic mappings can be increased by using phone-specific weights for the EMA coordinates of critical articulators. Likewise, context-dependent articulatory mappings could be used to minimize errors in EMA pellet positions that are critical to each phone, in this fashion improving synthesis quality and accent-conversion performance. Additional information on vocal tract geometry may also be used to improve synthesis performance. As an example, having access to the palate contour may be used to compute the distance (or contact) between passive and active articulators, or to extract tract constriction variables, which are known to have less variability than EMA pellet positions (McGowan, 1994; Mitra *et al.*, 2011).

The foreign accent conversion method described in this section uses GMM-based forward mapping. The GMM-based forward mapping is selected because of its accuracy and the flexibility to interpolate new sounds for the articulatory configuration not available in the training database. However, the GMM-based synthesizer uses the dynamics of estimated acoustic features to reduce temporal spectral discontinuities, hence, increasing the computational costs and latency during run-time. A few low-delay approximations are available but they are known to reduce the acoustic quality (Muramatsu *et al.*, 2008; Toda *et al.*, 2012). The method is thus unsuited for real-time conversion. In the next section, we describe a real-time articulatory synthesizer that exploits the temporal nature of speech in the articulatory feature. We also evaluate how effective is the synthesizer in foreign accent conversion.

6. ARTICULATORY-BASED CONVERSION OF FOREIGN ACCENTS WITH DEEP NEURAL NETWORKS[15]

In the previous section, we presented a GMM-based articulatory method for foreign accent conversion. The method was able to reduce the perceived non-native accents while preserving the voice-quality of the non-native speaker. However, the method suffers from a run-time inefficiency that involves the estimation of maximum-likelihood trajectories of acoustic features considering their dynamics in order to reduce the spectral discontinuities across adjacent frames. Such trajectory optimization is necessary to improve acoustic quality, but is computationally expensive as it requires the entire utterance to be processes at once, making the GMM-based approach inadequate for real-time accent conversion. Low-delay and low-latency implementations of the trajectory optimization process are available (Muramatsu *et al.*, 2008; Xingyu *et al.*, 2014), but only at the cost of reduced acoustic quality. In this section, we present a method that exploits the temporal nature of speech in articulatory input features to reduce discontinuities and avoid the expensive trajectory optimization of estimated acoustic features (output features), such that accent conversion is possible in real-time. The method utilizes deep neural networks (DNN) in modeling articulatory-acoustic mappings.

[15] The description of the method and the experimental results are reprinted with permission from "Data driven articulatory synthesis with deep neural networks," by Aryal and Gutierrez-Osuna, 2015(in press), *Computer Speech & Language,* ©2015 Elsevier B.V., and from "Articulatory-based conversion of foreign accents with deep neural networks," by Aryal and Gutierrez-Osuna, 2015, *Proceedings of INTERSPEECH*, pp. 3385-3389. ©2015 ISCA.

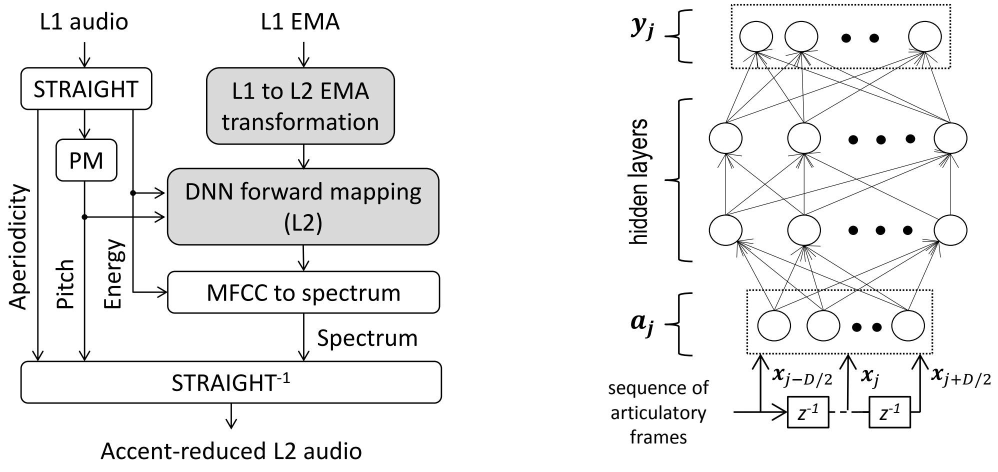
6.1 Deep neural network in articulatory-acoustic mappings

Non-parametric models such as neural networks have rarely been used in forward-mapping problems, where GMMs are considered the de-facto standard. One notable exception is the work by Kello et al. (2004), who used a single-layer multilayer perceptron (MLP) to estimate acoustic features (Fourier transform coefficients) from electromagnetic articulography (EMA), electropalatograph and laryngograph measurements. In an intelligibility test, the authors reported a word identification rate of 84% for synthesized speech, only 8% lower than that of the actual recordings.

Compared to single-layer MLPs, DNNs can be expected to provide higher forward-mapping accuracy. First, the presence of multiple hidden layers makes DNNs more flexible models, allowing them to represent complex functions with fewer hidden units. Second, DNNs are pre-trained as generative models in an unsupervised mode, a step that has been shown to guide the learning process towards parameters that support better generalization (Erhan *et al.*, 2010). These predictions have been corroborated in several speech-related applications (Hinton *et al.*, 2012; Uria *et al.*, 2012; Zen *et al.*, 2013), where DNN-based methods have surpassed the performance of state-of-the-art methods based on the HMM-GMM framework. Among these, a study on articulatory inversion by Uria et al. (Uria *et al.*, 2012) is particularly relevant here given the similarity between both problems. Using a DNN, the authors were able to estimate EMA pellet positions with an average root mean square error of 0.95mm on the MNGU0 test dataset, an error that was not only lower than that of a single-layer MLP but also the lowest among all previously published results on that dataset. A recent study by Andrew et al. (2013) on joint articulatory-acoustic modeling also highlights the superiority of deep learning techniques in this domain. The authors proposed a deep architecture for canonical correlation analysis (CCA) and tested it on the Wisconsin X-ray Microbeam Database (Westbury, 1994). Their deep CCA method achieved significantly higher correlation between the transformed acoustic and articulatory spaces than conventional CCA and kernel-based CCA (Arora and Livescu, 2013) and also compared favorably against kernel-CCA in terms of flexibility and the computational complexity. These results motivate our exploration of DNNs for real-time articulatory synthesis and its application in foreign accent conversion. In the following, we describe the proposed DNN-based method for real-time accent conversion.

6.2 DNN-based foreign accent conversion method

As shown in Figure 23a, the overall approach for foreign accent conversion consists of four main stages: (1) articulatory normalization to map L1 EMA positions into L2 articulatory space, (2) DNN forward mapping to estimate L2 acoustic parameters from normalized L1 EMA positions, (3) scaling of the L1 pitch contour to match the pitch range of the L2 speaker, and (4) reconstructing the speech waveform via STRAIGHT synthesis. The approach is similar to the GMM-based conversion in the previous section except for the DNN-based forward mapping, which we describe next. The other three phases have been already described in previous section (see section 5.1 for more detail).



(a) (b)

Figure 23: (a) DNN-based foreign accent conversion (PM: pitch modification) (b) Forward mapping using a DNN with a tapped-delay line input.

6.2.1 DNN-based forward mapping

Given a trajectory of articulatory features $\boldsymbol{x} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3} \dots \boldsymbol{x_T}]$ for an utterance, the DNN estimates the corresponding sequence of acoustic feature vectors $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2 \dots \boldsymbol{y}_T]$. As illustrated in Figure 23b, the DNN consists of an input layer, an output layer, and multiple layers of hidden units between them. In this particular topology, units in a layer are fully connected to units in the immediate layer above it, but there is no connection among units within a layer. The network contains a tapped-delay line to contextualize the input with features from past and future frames, resulting in the input vector $\boldsymbol{a}_j = \{\boldsymbol{x}_{j-D/2} \dots \boldsymbol{x}_j \dots \boldsymbol{x}_{j+D/2}\}$, where $\boldsymbol{x}_j$ is the articulatory configuration at frame j, and D is the number of delay units. The DNN consists of Gaussian input units and binary hidden units, all units with sigmoid activation function.

Training the DNN is a two stage process. First, a Gaussian-Bernoulli Boltzmann machine (Cho *et al.*, 2013) is trained in an unsupervised fashion. Finally, a layer of output nodes (one node for each acoustic parameter) is added on top of the trained GDBM to form a DNN, which is then fine-tuned via back-propagation (Rumelhart *et al.*, 1986). See Appendix A for more detail.

6.2.2 Global variance adjustment

Statistical mappings are known to over-smooth the acoustic trajectories resulting in muffled sounds (Toda *et al.*, 2007). For this reason, our GMM-based accent conversion method incorporated the global variance (GV) of the acoustic feature vectors to reduce the over-smoothing effects. To ensure a fair comparison with the GMM-based method, we adjust the DNN estimated acoustic features as follows. Let the acoustic feature vector estimated by the DNN at frame j of the test utterance be $\boldsymbol{y}_j$, then, the GV-adjusted feature vector $\hat{\boldsymbol{y}}_j$ is given by:

$$\hat{\boldsymbol{y}}_j = \left(\boldsymbol{y}_j - \boldsymbol{\mu}\right)\mathbf{A} + \boldsymbol{\mu} \tag{24}$$

where $\boldsymbol{\mu}$ is the mean of the estimated acoustic feature vectors, and $\boldsymbol{A}$ is a diagonal matrix whose elements are the square roots of the ratios between the GVs for the natural and estimated trajectories. Calculating the exact values for $\boldsymbol{\mu}$ and $\mathbf{A}$ requires the estimated acoustic features for the entire utterance, which is not possible in real-time conversion. Therefore, we calculate these parameters $(\boldsymbol{\mu}, \mathbf{A})$ for all the training sentences and use their average value as an approximation during run-time.

6.3 Performance of DNN-based forward mapping

Before evaluating the performance of DNN forward mapping in accent conversion, we set out to evaluate how effective the DNNs are in articulatory-to-acoustic mapping. For this purpose, we compared the proposed DNN-based mapping method against two GMM-based methods based on Toda *et al.* (2004). Since the GV adjustments reduce the mapping accuracy which would distort results from the objective tests[16], the DNN-based and GMM-based methods in this comparison do not incorporate GV.

6.3.1 GMM-based baseline methods

The first method, which we denote by sGMM, ignores dynamic information and serves as a baseline for real-time synthesis. Namely, sGMM performs a frame-by-frame mapping from articulatory positions onto *static* acoustic features (MFCCs). The second method, dGMM, incorporates the dynamics of acoustic features to improve the forward-mapping accuracy. Namely, dGMM predicts not only MFCCs but also delta-MFCCs, and then performs the computationally-intensive trajectory optimization (Toda *et al.*, 2004). As such, dGMM is unsuited for real-time synthesis so it should be taken as an upper bound on accuracy.

The GMMs required for both methods are trained to model the joint distribution of articulatory and acoustic features $\boldsymbol{Z_t} = [\boldsymbol{x}_t, \boldsymbol{Y}_t]$, where $\boldsymbol{x_t}$ is the articulatory feature

[16] The foreign accent conversion methods are evaluated through subjective listening tests. Thus, incorporating the global variance in their mapping methods does not distort the comparison results.

vector and $\boldsymbol{Y}_t = [\boldsymbol{y}_t,\ \Delta\boldsymbol{y}_t]$ is an acoustic feature vector containing both static and delta values at frame t. The joint distribution is given by:

$$p\left(\boldsymbol{Z}_t \middle| \boldsymbol{\lambda}^{(z)}\right) = \sum_{m=1}^{M} \alpha_m\, \mathcal{N}(\boldsymbol{Z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}) \tag{25}$$

where α_m is the scalar weight of the m^{th} mixture component and $\mathcal{N}(; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)})$ is the Gaussian distribution with mean $\boldsymbol{\mu}_m^{(z)}$and covariance matrix $\boldsymbol{\Sigma}_m^{(z)}$:

$$\boldsymbol{\mu}_m^{(Z)} = \left[\boldsymbol{\mu}_m^{(x)}\ \boldsymbol{\mu}_m^{(Y)}\right], \quad \boldsymbol{\Sigma}_m^{(Z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xY)} \\ \boldsymbol{\Sigma}_m^{(Yx)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \tag{26}$$

In what follows, we use the symbol $\boldsymbol{\lambda}^{(z)} = \{\alpha_m, \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}\}$ to denote the full parameter set for the GMM. Given a trained GMM and a test sequence of articulatory feature vectors $\boldsymbol{x} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3} \ldots \boldsymbol{x_T}]$, we generate separate predictions of acoustic feature vectors $\boldsymbol{y} = [\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots \boldsymbol{y_T}\ ]$ for the two GMM variants as follows:

1. For sGMM, we ignore the acoustics dynamics and calculate the static acoustic feature vector at frame t as the minimum mean square error (MMSE) estimate:

$$\widehat{\boldsymbol{y}}_{t,MMSE} = \sum\nolimits_{m=1}^{M} P\left(m \middle| \boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}\right) \boldsymbol{E}_{m,t}^{(y)} \tag{27}$$

   where $\boldsymbol{E}_{m,t}^{(y)}$ is the subset of static features in the conditional expected value $\boldsymbol{E}_{m,t}^{(Y)}$, as given by

$$\boldsymbol{E}_{m,t}^{(Y)} = \boldsymbol{\mu}_m^{(Y)} + (\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}) \boldsymbol{\Sigma}_m^{xx^{-1}} \boldsymbol{\Sigma}_m^{(xY)}. \tag{28}$$

2. For dGMM, we calculate the maximum likelihood estimate of the acoustic trajectory considering the dynamics, as given by:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}} \; P\big(\boldsymbol{Y}\big|\boldsymbol{x}, \boldsymbol{\lambda}^{(z)}\big) \tag{29}$$

where $\boldsymbol{Y} = [\boldsymbol{y_1}, \Delta\boldsymbol{y_1}, \boldsymbol{y_2}, \Delta\boldsymbol{y_2}, \ldots \boldsymbol{y_T}, \Delta\boldsymbol{y_T}\ ]$ is the time sequence of acoustic vectors (both static and dynamic). We solve for $\hat{\boldsymbol{y}}$ in equation (23) iteratively via the EM algorithm; see (Toda *et al.*, 2004) for more details.

6.3.2 Experimental

We evaluated the three forward mappings (DNN, sGMM, dGMM) on the corpus described earlier in section 5.2. The corpus contained simultaneous recordings of acoustics and articulatory trajectories recorded via electromagnetic articulography (EMA) from a native and a non-native speaker of American English. Out of the two speakers, we used the native speaker of American English to avoid effects of inconsistencies in non-native productions in the evaluation. Out of the 344 sentences recorded, 294 randomly-selected sentences were used to train the model and the remaining 50 sentences were used only for test synthesis. As explained in section 5.2, we extracted articulatory and acoustic features for all the utterances in the corpus. For each frame, the articulatory feature vector consisted of 15 parameters: the anteroposterior and superoinferior coordinate of six EMA pellets, pitch ($\log f_0$), loudness ($MFCC_0$) and *nasality*; the acoustic feature vector consists of acoustic parameters ($MFCC_{1-24}$). All the acoustic and articulatory parameters were normalized to zero-mean and unit-variance.

For the two GMM-based mappings, we trained GMMs with 128 mixture components on the joint distribution of articulatory and acoustic features (including delta) using the Netlab toolbox (Nabney, 2002). Once the GMMs were trained, we estimated acoustic features using *sGMM* and *dGMM* methods as described by equations (27) and (23), respectively. For the DNN mapping, we used a tapped-delay line with delay units of 10 ms (≈2 frames), and evaluated tapped-delays with 2, 4, 6, and 8 delay units. As an example, for a delay line with 6 units the input vector contains features from 7 frames covering 60 ms of articulatory context (30ms backward, 30 ms forward). DNNs were implemented using the Deepmat toolbox (Cho, 2013).

Once a vector of MFCCs was predicted by either of the three mappings (DNN, sGMM, dGMM), we used the STRAIGHT synthesis engine to generate the waveform using the estimated spectral envelope, and the signal aperiodicity and pitch. The overall process is illustrated in the figure below.

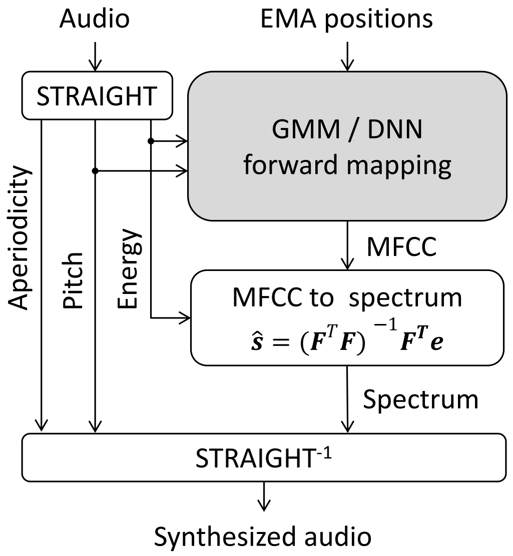

Audio
EMA positions
STRAIGHT
GMM / DNN forward mapping
Aperiodicity
Pitch
Energy
MFCC
MFCC to spectrum
$\hat{\boldsymbol{s}} = (\boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T\boldsymbol{e}$
Spectrum
STRAIGHT^{-1}
Synthesized audio

Figure 24. Signal processing flow during articulatory synthesis.

Following Toda *et al.* (2004), we evaluated the forward mappings based on the Mel-Cepstral distortion between ground-truth and estimated acoustic features:

$$MCD = \frac{10}{\ln 10}\sqrt{2\sum\nolimits_{d=1}^{24}\left(\boldsymbol{y}_t^{(d)} - \hat{\boldsymbol{y}}_t^{(d)}\right)^2} \tag{30}$$

where $\hat{\boldsymbol{y}}_t^{(d)}$ is the d^{th} component of the estimated acoustic feature vector (i.e., MFCC) at the t-th frame in a test utterance, and $\boldsymbol{y}_t^{(d)}$ is the ground-truth value extracted from the acoustic recording. MCDs were calculated only on non-silent frames.

Acoustic predictions for the three forward mappings on a typical test utterance are illustrated in Figure 25 alongside the ground truth. Because of the large number of inputs and outputs, we have only included trajectories for three articulatory coordinates (TT_x, TT_y, and LL_y) and one acoustic feature ($MFCC_2$). Predictions from the sGMM display a number of unnatural transitions or glitches (see arrows in the figure), which are perceptible and have a detrimental effect on synthesis quality. Although the dGMM avoids such unnatural transitions by accounting for the dynamics of acoustic features in the trajectory optimization stage, it suffers from over-smoothing[17] effects, which are also perceptible and also clearly seen in the figure. By comparison, predictions from the DNN follow the target trajectory closely without introducing discontinuities in the derivative or over-smoothing.

[17] A method known as global variance (Toda *et al.*, 2007) has been suggested as a solution to the over-smoothing problem in the dGMM. However, the global variance method also increases prediction errors, so was not considered in this study as it would distort results from the objective tests.

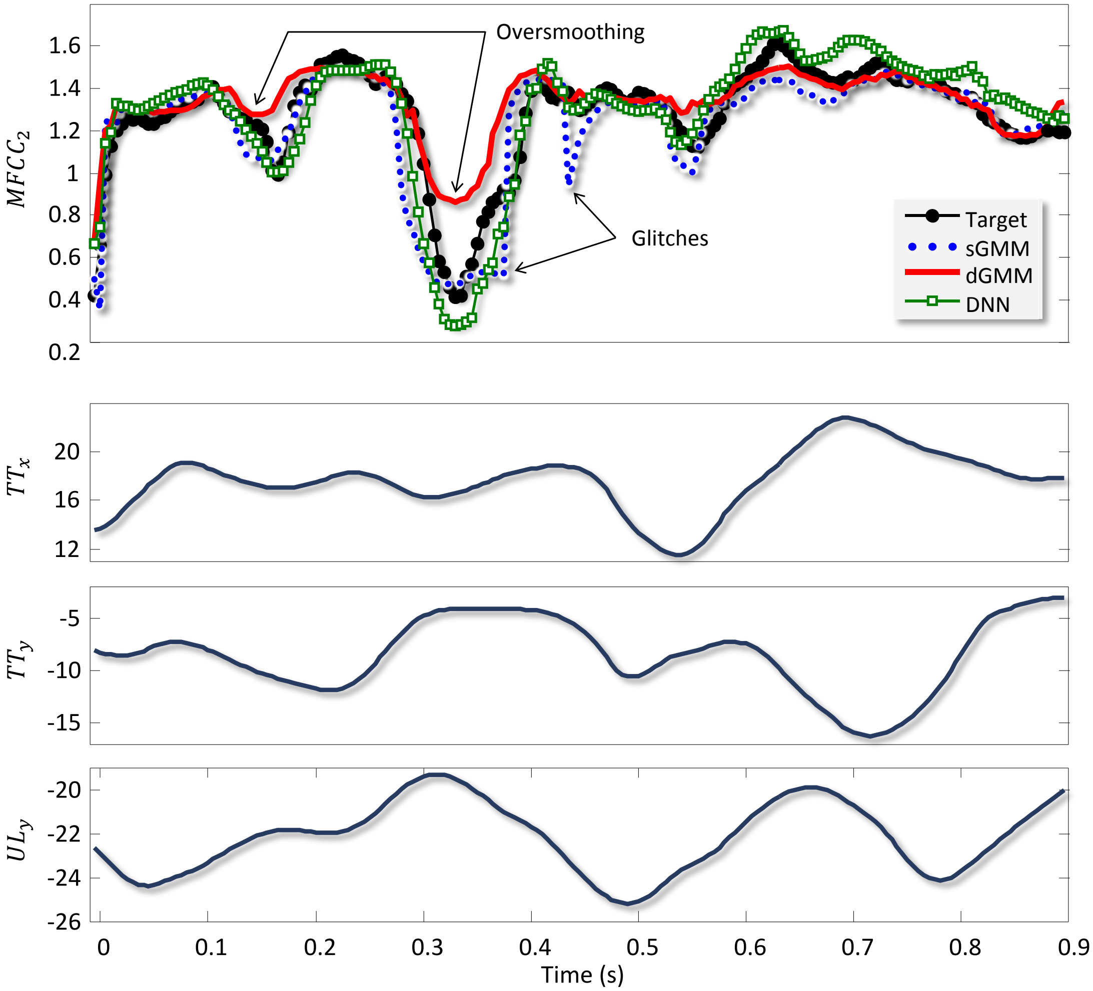


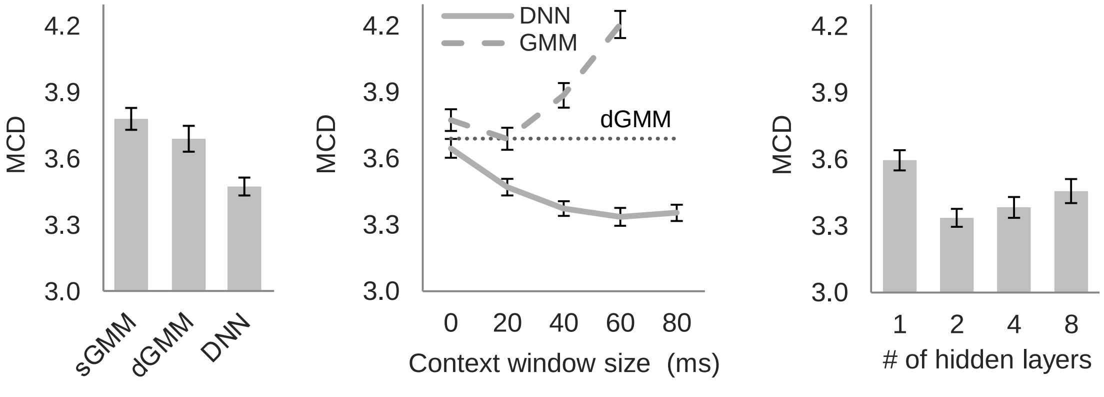
Figure 25: Trajectories of selected acoustic and articulatory features from a typical test utterance. The top plot shows the second MFCC predicted by the DNN, sGMM and dGMM alongside the target trajectory extracted from the audio recording of the same sentence. The bottom plots show the trajectories of a few articulatory input features for the same utterance. TT_x: anteroposterior position of the tongue tip, TT_y: height of the tongue tip, LL_y: height of the upper lip.

We evaluated the forward mappings through a series of objective and subjective tests. In a first experiment, we compared the DNN against the two GMM mappings (sDNN, dGMM) in terms of their mapping accuracy (Mel-Cepstral distortion). Next, we evaluated the effect of tapped-delay length (experiment 2) and network depth (experiment 3) on Mel-Cepstral distortion, followed by a comparison of synthesis-time (experiment 4). In a final experiment, we compared the best performing DNN and GMM through a perceptual listening test.

6.3.3 Experiment 1: Comparison of DNN vs. GMM

In the first experiment, we compared the accuracy of the DNN forward mapping against the two reference GMM methods. The DNN had a tapped-delay line with 2 delay units (a context window size of 20 ms) and two hidden layers of 512 units each. This simple architecture was selected to keep the number of model parameters comparable to that of the GMMs.

Figure 26a summarizes the average MCDs of the three methods. The dGMM and DNN models achieve lower Mel-Cepstral distortion than the sGMM mapping. This is consistent with findings from previous studies (Toda *et al.*, 2004; Nakamura *et al.*, 2006), and shows that exploiting temporal information (as done by the dGMM and DNN) provides higher accuracy than a frame-by-frame mapping (sGMM),. More importantly, the DNN reduces Mel-Cepstral distortion by 6% compared to the dGMM ($p < 0.001$, pairwise t-test), indicating that comparable (if not better) accuracy can be achieved at a fraction of the synthesis time required by the dGMM.



(a) (b) (c)

Figure 26: (a) Experiment 1: Mel cepstral distortion (MCD) for the DNN, sGMM and dGMM mappings. (b) Experiment 2: MCD for the DNN and GMM as a function of the input articulatory context window. (c) Experiment 3: MCD for the DNN as a function of the number of hidden layers; error bars denote standard errors of means.

6.3.4 Experiment 2: Context length

In the second experiment, we trained DNNs with tapped-delay line lengths of 0, 2, 4, 6 and 8 units, corresponding to temporal window sizes of 0, 20, 40, 60 and 80 ms, respectively. In each of these DNNs, we kept the same number of hidden layers and hidden units used in the first experiment. Figure 26b summarizes results in terms of the Mel-Cepstral distortion, including that of the dGMM as a reference. Regardless of context length, the DNNs result in lower Mel-Cepstral distortion than the dGMM, the difference being statistically significant except for a context window size of 0 ms (i.e., a frame-by-frame mapping). More importantly, the Mel-Cepstral distortion decreases as the context window size increases, reaching a minimum with a 60 ms context window –a 9.8% reduction compared to the dGMM.

As part of this experiment we also sought to answer whether the same improvements in performance could be achieved by a GMM with a tapped-delay line. For this purpose we trained four GMMs with tapped-delay lines of 0, 20, 40 and 60 ms, respectively. Results are shown in Figure 26b; GMM mappings had higher Mel-Cepstral distortion than the corresponding DNN regardless of context window size. More importantly, whereas the DNN is able to take advantage of the added information in the tapped-delay line (up to 60 ms), the GMM accuracy decreases markedly for context window sizes larger than 20ms. This result may be explained by the fact that the tapped-delay features tend to be highly correlated, which may lead to near-singular covariance matrices in the GMM.

6.3.5 Experiment 3: Network depth

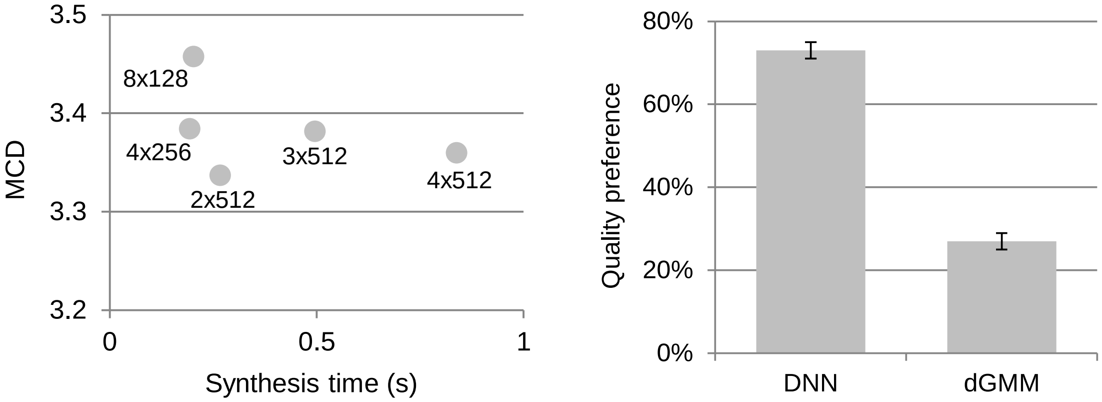
In the third experiment, we sought to determine whether the complexity of the forward mapping justifies the use of a DNN; a DNN can model complex nonlinear functions with fewer parameters than a single-hidden MLP, but requires considerably longer training times. To answer this question, we trained four models: a single-layer MLP with 1024 hidden nodes, and three DNNs with 2, 4 and 8 hidden layers; the numbers of hidden units per layer in the DNN were adjusted so that the total number of hidden units remained constant across models (i.e. 1024). The tapped-delay line was fixed to 60 ms, the optimal context length found in the previous experiment. The MLP was trained using standard back-propagation (Rumelhart *et al.*, 1986).

Figure 26c summarizes the average Mel-Cepstral distortion for the four architectures; the three DNNs outperformed the MLP (pairwise t-test $p < 0.01$), which suggests that a single-layer network is insufficient to model the articulatory-to-acoustic mapping. The minimum Mel-Cepstral distortion —a 7% reduction compared to a single-layer MLP, was obtained for a DNN with 2 hidden layers.

6.3.6 Experiment 4: Synthesis time

In the fourth experiment, we compared the synthesis time of the DNN and dGMM mappings. Both models were run on a Windows 7 Enterprise machine with an Intel Core i7-2600 @3.4 GHz processor; models were implemented and run under Matlab v.7.14.

On average, the dGMM method required 39 seconds of synthesis time for each second of speech, rendering it unsuited for real-time synthesis (results not shown). In the case of the DNN, synthesis time depended on the network size, but increased linearly with the number of connections in the network. Figure 27a shows the relationship between Mel-Cepstral distortion and synthesis time for five DNN structures, three from the third experiment (2×512, 4×256 and 8×128 hidden units, 60 ms context) and two relatively larger networks (3×512 and 4×512 hidden units) trained specifically for this experiment. The largest among them, a DNN with 4 layers of 512 hidden units, required 838 ms for each second of speech, suitable for real-time synthesis. Smaller networks are even more efficient: a DNN with 2 layers of 512 hidden units required only 267 ms for each second of speech, and achieved the lowest Mel-Cepstral distortion.



(a) (b)

Figure 27: (a) Experiment 4: Synthesis time of a DNN mapping increases with the size of the network. (b) Experiment 5: Pairwise comparison between DNN and dGMM synthesis; error bars denote standard errors of means.

6.3.7 Experiment 5: Subjective assessment

In the final experiment, we evaluated the best-performing DNN (2×512 hidden units and 60 ms context window) against the conventional dGMM of Toda et al. (Toda *et al.*, 2004) through a listening test. Our goal was to determine whether the improvement in Mel-Cepstral distortion achieved by the DNN (a reduction of 9.8%) was perceptually significant.

For the subjective listening test, we recruited participants through Mechanical Turk, Amazon's online crowdsourcing tool. Participants listened to pairs of synthesis of the same sentence (one from the DNN, another from the dGMM) and were asked to select the utterance with the best quality in terms of naturalness, distortion, and intelligibility. 30 listeners participated in this test, each participant rating 30 pairs of utterances. Order of presentation within a pair (DNN vs. dGMM) was randomized to avoid order bias. Shown in Figure 27b, DNN syntheses were rated as more natural than dGMM syntheses in 73% of the cases, which is significantly higher than 50% chance level (pairwise t-test, $p < 0.001$). This result corroborates the objective comparisons, and indicates that the DNN mapping can synthesize utterances of higher perceptual quality than the conventional dGMM.

6.3.8 Discussions on the performance of DNN-based forward mapping

We have presented a real-time articulatory synthesis method that exploits dynamic information in the articulatory trajectories to increase the accuracy of the forward mapping. Namely, our approach uses a tapped-delay line to concatenate articulatory feature vectors (EMA positions) from nearby frames, and a DNN to map the concatenated articulatory input vector into the corresponding acoustic observations (MFCCs). We compared the DNN against two GMM-based articulatory synthesizers, one that performs a frame-by-frame mapping (sGMM) and one that also incorporates speech dynamics (dGMM) as proposed by Toda et al. (2004). As our results show, the DNN is able to take advantage of the additional information in the articulatory tapped-delay line while keeping synthesis time below frame rate, surpassing the accuracy of both GMM-based methods through objective evaluations (Mel Cepstral distortion) and the subjective quality of the dGMM through listening tests.

Though GMMs are easier to train than DNNs, our results show they are unable to exploit the added temporal information via a tapped-delay line. This is partly due to the fact that the number of model parameters in a GMM increases quadratically with the number of input features, which can lead to over-fitting given the limited amount of training data. More importantly, tapped-delay features are likely to be correlated since they are time-delayed versions of the same signal, which may lead to near-singular covariance matrices in the GMMs. Though linear dimensionality reduction techniques (e.g., principal components analysis) may be used to decorrelate the input features, research in speech recognition (Bao *et al.*, 2012) indicates that such techniques cannot compete with the capabilities of DNNs.

The dGMM and DNN articulatory synthesizers represent two distinct alternatives to incorporate speech dynamics. dGMMs can be trained relatively fast, but have long synthesis times due to the trajectory optimization post-processing stage; in our experiments, each second of speech required an average of 39 seconds of synthesis time on a contemporary desktop computer. By contrast, training a DNN is time consuming, but this is usually a one-time process that can be done offline. Once trained, the DNN has a short synthesis time[18] (e.g., 267 ms for our best-performing DNN). This makes the DNN ideally suited for other real-time applications of articulatory synthesis such as silent speech interfaces (Denby *et al.*, 2010).

After establishing the performance of DNN in forward mapping (within speaker), next, we examine whether the DNN articulatory synthesizer can also outperform the GMM articulatory synthesizer *across speakers*, as needed for accent conversion.

[18] Although the DNN uses a tapped-delay line that extends 30 ms into the future, this latency time (< 200 ms) is considered acceptable for real-time communication (ITU-T, 2003).

6.4 Evaluation of foreign accent conversion with DNN

We evaluated the DNN and GMM accent conversion models on an experimental corpus of parallel recordings of articulatory and audio signal from a native and a non-native speaker of American English (Felps *et al.*, 2012) collected via Electromagnetic articulography (EMA). Both speakers recorded the same set of 344 sentences, out of which 294 sentences were used for training the model and the remaining 50 sentences were used only for testing. See section 5.2 for more detail on the corpus, processing and feature extraction.

The baseline GMMs were trained with 128 mixture components (full covariance), whereas the DNNs contained 2 layers of 512 hidden nodes, and a 60ms tapped-delay input (seven 10-ms frames: 3 previous, 1 current, 3 future). We have found these GMM and DNN structures to perform reliably in forward mapping tasks (sections 5.3 and 0).

In order to evaluate the DNN-based accent conversion method, we synthesized test sentences in five experimental conditions –see Table 6: (i) the proposed accent conversion method (AC_{DNN}), (ii) articulatory resynthesis by driving the DNN with L2 articulators ($L2_{EMA}$), (iii) accent conversion using the GMM-based method, as described in section 5.1 (AC_{GMM}), (iv) MFCC compression of L2 speech ($L2_{MFCC}$), and (v) L1 utterances modified to match the vocal tract length (Sundermann *et al.*, 2003) and pitch range of L2 ($L1_{GUISE}$). We evaluated these conditions through a series of subjective listening tests on Mturk, Amazon's crowd sourcing tool. To qualify for the study, participants were required to reside in the United States and pass a screening test that consisted of identifying various American English accents, including Northeast, Southern, and General American.

Table 6: Experimental conditions for the listening tests.

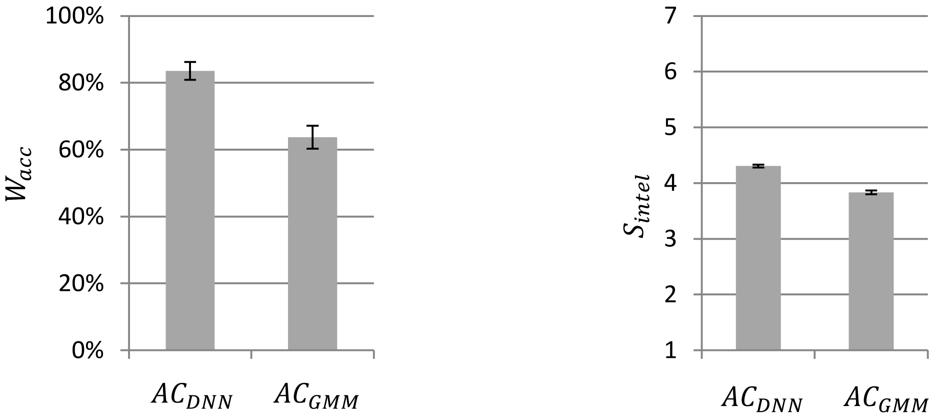
Experimental conditions	Aperiodicity and energy	Pitch	Articulators	Spectrum	Forward-mapping model
AC_{DNN}	L1	L1 scaled to L2	L1 mapped to L2	L2 forward mapping	DNN
$L2_{EMA}$	L2	L2	L2	L2 forward mapping	DNN
AC_{GMM}	L1	L1 scaled to L2	L1 mapped to L2	L2 forward mapping	GMM
$L2_{MFCC}$	L2	L2	N/A	L2 MFCC	N/A
$L1_{GUISE}$	L1	L1 scaled to L2	N/A	L1 warped to L2	N/A

6.5 Results

6.5.1 Intelligibility assessment

In a first listening test we assessed the intelligibility of the proposed method (AC_{DNN}). We asked a group of participants (N=15) to transcribe 46 test utterances from AC_{DNN}, and also rate the (subjective) intelligibility (S_{intel}) of those utterances using a seven-point Likert scale (1: not intelligible at all, 3: somewhat intelligible, 5: quite a bit intelligible, and 7: extremely intelligible). From the transcription, we calculated word accuracy (W_{acc}) as the ratio of the number of correctly-identified words to the total number of words in the utterance. To compare the intelligibility of the proposed method against the baseline method, we used the same set of test sentences in section 5.3.2.

Figure 28 shows the word accuracy and the subjective intelligibility ratings for the two accent-conversion models (AC_{DNN} and AC_{GMM}). The DNN model had higher scores ($W_{acc} =$ 84%, $S_{intel} =$ 4.3) than the baseline GMM model ($W_{acc} =$ 64%, $S_{intel} =$ 3.84), and the differences were statistically significant (W_{acc}: $t(45) = 7.4, p < 0.001$; S_{intel}: $t(45) = 3.66, p < 0.001$).

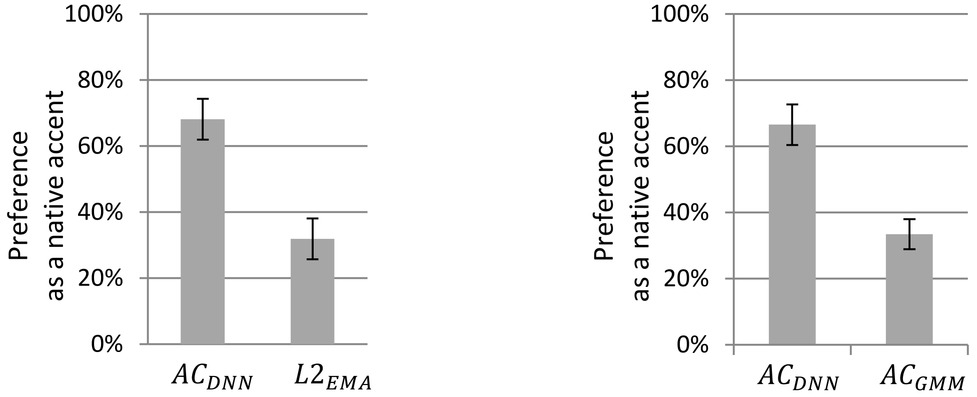


(a) (b)

Figure 28: (a) Word accuracy and (b) subjective intelligibility ratings for AC_{DNN} and AC_{GMM}.

6.5.2 Assessment of non-native accentedness

In a second set of listening tests, we examined the ability of the DNN to reduce the perceived non-native accent of L2 utterances. Participants were asked to listen to pairs of utterances –one from the accent conversion (AC_{DNN}) method, the other an articulatory resynthesis of the L2 utterance ($L2_{EMA}$) for the same sentence, and select the most native-like. The articulatory resynthesis ($L2_{EMA}$) was used instead of the original L2 recording to account for losses in acoustic quality due to the articulatory-synthesis step in the accent conversion process, which are known to affect accent perception (Felps *et al.*, 2009). As before, we tested on the same subset of 15 test sentences from section 5.3.3 so that the results could be compared.



(a) (b)

Figure 29: Subjective evaluation of accentedness. Participants selected the most native-like utterances (a) between AC_{DNN} vs. L2 articulatory resynthesis, and (b) between AC_{DNN} vs. AC_{GMM}.

Participants listened to 30 pairs of utterances (15 $AC_{DNN} - L2_{EMA}$ pairs and 15 $L2_{EMA} - AC_{DNN}$ pairs) presented in random order to account for ordering effects. As shown in Figure 29a, participants rated AC_{DNN} more native-like than L2 articulatory resynthesis in 68% ($s.e = 6\%$) of the sentences, which is significantly higher ($t(14) = 3.03, p < 0.01$) than the 50% chance level. *This result shows that the proposed DNN-based method is effective in reducing perceived non-native accents.*

Next, we compared the DNN accent conversion method against the baseline GMM method. For this purpose, a different group of participants listened to the 30 pairs of utterances (15 $AC_{DNN} - AC_{GMM}$ pairs and 15 $AC_{GMM} - AC_{DNN}$ pairs) presented in random order. Shown in Figure 29b, AC_{DNN} utterances were rated as more native-like than AC_{GMM} utterances in 67% ($s.e. =$ 5%) of the sentences, which is also significantly higher than the 50% chance level ($t(14) = 3.6674, p < 0.01$).

6.5.3 Voice identity assessment

In a third and final listening experiment we evaluated if the DNN accent-conversion method was able to preserve the voice identity of the L2 speaker. For this purpose, participants were asked to compare the voice similarity between pairs of utterances, one from AC_{DNN}, the other from $L2_{MFCC}$ (MFCC compression of the original L2 recordings). As a sanity check, we also included pairs of utterances from $L2_{MFCC}$ and $L1_{GUISE}$, the latter a simple guise of L1 utterances to match the pitch range and vocal tract length of the L2 speaker. The utterances in each pair were linguistically different, and presentation order was randomized for conditions within each pair and for pairs of conditions. Participants ($N = 15$) rated 40 pairs, 20 from each group ($L2_{MFCC} - AC_{DNN}$, $L2_{MFCC} - L1_{GUISE}$) randomly interleaved, and were asked to (1) determine if the utterances were from the same or a different speaker and (2) rate how confident they were in their assessment using a seven-point Likert scale (1: not confident at all, 3: somewhat confident, 5: quite a bit confident, and 7: extremely confident). The responses and their confidence ratings were then combined to form a voice similarity score (VSS)

ranging from -7 (extremely confident they are different speaker) to $+7$ (extremely confident they are from the same speaker).

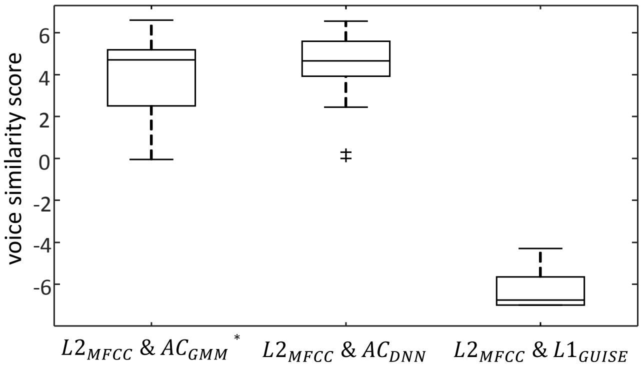


Figure 30: Average pairwise voice similarity scores. $L2_{MFCC}$& AC_{GMM} * from section 5.3.4.

Figure 30 shows the boxplot of average VSS between the pairs of experimental conditions. Participants were 'quite' confident ($VSS = 4.3, s.e. = 0.5$) that the $L2_{MFCC}$ and AC_{DNN} were from the same speaker, suggesting that the method successfully preserved the voice-identity of L2 speaker. The VSS was also comparable ($t(14) = -0.37, p = 0.71$) to the VSS between AC_{GMM} and $L2_{MFCC}$ ($VSS = 4.0, s.e. = 0.5$) reported for the baseline GMM method in previous section. The participants were also 'quite' confident that ($VSS = -6.3, s.e. = 0.2$) the $L2_{MFCC}$ and $L1_{GUISE}$ were from different speakers, corroborating our prior finding (section 5.3.4) that a simple guise of L1 utterances is not sufficient to match the voice of the L2 speaker. These findings suggest that the run-time capabilities of the DNN did not compromise its ability to preserve the voice identity.

6.6 Conclusion

We have presented an articulatory method for real-time modification of non-native accents. The approach uses a DNN with a 60ms tapped-delay input to map L2 articulatory trajectories into L2 acoustic observations (MFCCs). Driving the DNN with articulatory trajectories from an L1 speaker—normalized to the L2 articulatory space—results in speech that captures the linguistic gestures of the L1 speaker and the voice quality of the L2 speaker.

We evaluated the DNN accent-conversion method against the baseline GMM method. Accent conversions with the DNN were more intelligible and were perceived as more native-like than those using the GMM. A possible explanation for the difference in perceived accentedness between both methods is that acoustic quality affects the perception of non-native accents (i.e., the lower the quality, the higher the non-native rating) (Felps *et al.*, 2009); although both methods use articulatory synthesis, the comparison in section 6.3.7 above shows that the DNN tends to synthesize speech of higher acoustic quality than the GMM.

7. ACOUSTIC VS. ARTICULATORY-BASED STRATEGIES

In the previous section, we showed how a statistical parametric articulatory synthesizer can be driven by the native articulators to generate speech with native like accent but the voice of the non-native speaker. Our focus on articulatory-strategies for foreign accent conversion in this dissertation work had two main motivations. First, the voice-independent representation of linguistic gestures via articulatory data facilitates transferring accents from one speaker to another without affecting the voice-quality. Secondly, the articulatory-based approach has a theoretical basis on the modulation theory of speech (Traunmüller, 1994), in which the articulatory synthesizer for the non-native speaker acts as the voice quality carrier, and the articulatory data from a native speaker modulates the synthesizer generating speech with native linguistic gestures. However, the current technologies to collect articulatory data such as EMA, X-ray Microbeam, MRI are not only expensive and invasive but also limited to the laboratory setting. In contrast, our acoustic-based strategy is cost effective and more practical since it uses audio recordings only. In addition, the acoustic-based method using cross-speaker statistical mapping was also found effective in reducing the perceived non-native accents while preserving the voice-quality of the non-native speaker. Given the accessibility of the acoustic-based strategy, here, we set out to compare its performance in reducing the perceived non-native accents against the theoretically-sound articulatory-based strategy.

7.1 Comparison between the articulatory and acoustic-based strategies

To determine whether the accent conversion is more effective in the acoustic space or in the articulatory space, we compare the two statistical methods presented in this work: for acoustic-based strategy, we choose the GMM-based spectral mapping method as described in section 4.1, and for the articulatory-based strategy, we choose the method described in section 5.1. Both methods use GMMs to model the joint distribution of the input and output features, and estimate the maximum likelihood of trajectories of acoustic parameters considering their dynamics and the global variance. Despite the similarity in the models, direct comparison of accent conversions from these two methods is not possible because of the difference in their synthesis quality. As discussed earlier, differences in acoustic quality are known to interact with the perception of non-native accents (Felps *et al.*, 2009).

The main reason behind the differences in the acoustic quality between the two methods is inherent to the synthesizers used in these methods. In the articulatory-based method, the synthesis is driven using articulatory data (six fleshpoint trajectories captured via EMA) from the reference native utterance; whereas the synthesizer in acoustic-based method is driven by the acoustic features (MFCCs in our case). EMA being less informative of the phonetic information that the acoustic features, the quality of EMA driven synthesizer in articulatory-based method is lower than the acoustic driven synthesizer of acoustic-based method. In order to account for differences in acoustic quality between the two methods, in this study, we build an *equivalent articulatory synthesizer* for the acoustic-based accent conversion method.

7.2 Equivalent articulatory synthesizer for the acoustic-based strategy

As described in section 4.1, given the sequence of acoustic feature vectors $(\boldsymbol{y}_{L1})$ from an utterance of the reference native speaker (L1), the acoustic-based conversion method estimates the trajectories of acoustic feature vectors $(\boldsymbol{y}_{L2})$ for the non-native speaker (L2) using a GMM-based cross-speaker spectral mapping $(g: \boldsymbol{y}_{L1} \rightarrow \boldsymbol{y}_{L2})$. The objective of the *equivalent articulatory synthesizer* is to have the same effect of segmental modification caused by the cross-speaker spectral mapping $(g: \boldsymbol{y}_{L1} \rightarrow \boldsymbol{y}_{L2})$ but using the L1 articulatory features as the input features from the same utterance instead of the acoustic features. In other words, we seek to build a cross-speaker forward mapping $(f: \boldsymbol{x}_{L1} \rightarrow \boldsymbol{y}_{L2})$ such that for a given L1 utterance (with the sequence of articulatory features, $\boldsymbol{x}_{L1}$ and the sequence of acoustic features, $\boldsymbol{y}_{L1}$), the estimated sequence of L2 acoustic feature vectors, $f(\boldsymbol{x}_{L1})$ is the equivalent to the one given by the cross-speaker spectral mapping $g(\boldsymbol{y}_{L1})$ —see Figure 31. In the following, we describe a method to build such cross-speaker forward mapping.

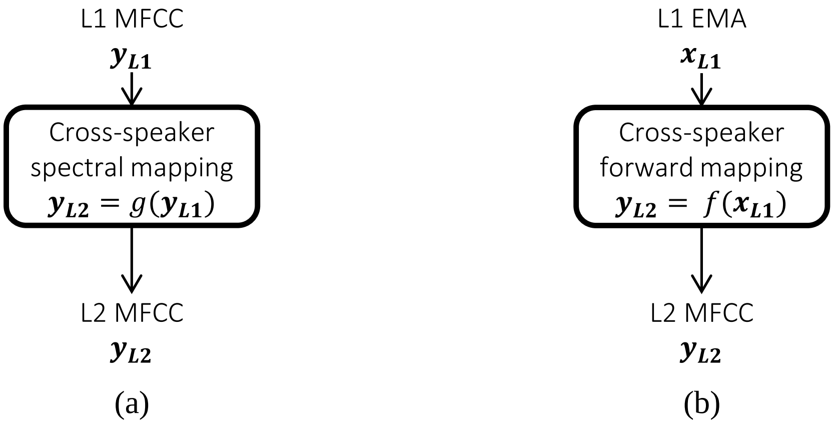


Figure 31: (a) cross-speaker spectral mapping for acoustic-based accent conversion, and (b) a cross-speaker forward mapping for the equivalent articulatory synthesis of acoustic-based accent conversion.

7.2.1 Training the cross-speaker forward mapping

A two-step process for training the cross-speaker forward mapping (f) is shown in Figure 32. In the first step, we estimate the L2 acoustic features for each L1 utterance in the training set using the cross-speaker spectral mapping function $(g: \boldsymbol{y}_{L1} \rightarrow \boldsymbol{y}_{L2})$ of the acoustic-based method. Note that, the resulting sequence of estimated acoustic feature vectors $g(\boldsymbol{y}_{L1})$ for each training sentence has the linguistic gestures of the reference L1 utterance but the voice-quality of the L2 speaker. In the second step, we build the GMM-based cross-speaker forward mapping by training it on the joint distribution of the L1 articulatory features $\boldsymbol{x}_{L1}$ and the estimated L2 acoustic features $g(\boldsymbol{y}_{L1})$ for the same.

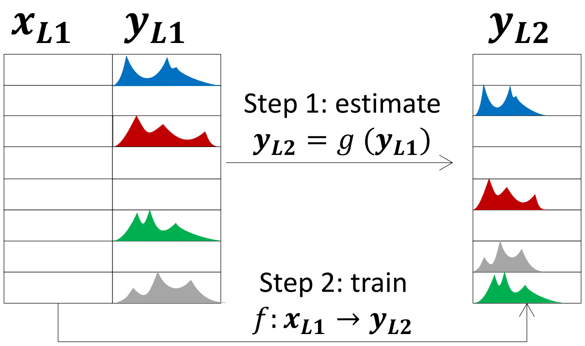


Figure 32: Two-step process for building a cross-speaker forward mapping, $f: \boldsymbol{x}_{L1} \rightarrow \boldsymbol{y}_{L2}$.

Once the cross-speaker forward mapping is available, for a given test utterance from the native speaker, we can estimate a sequence of equivalent L2 acoustic feature vectors. We then convert the estimated acoustic feature vectors into the waveform using the STRAIGHT synthesis engine.

7.3 Experimental validation

We performed a series of subjective listening tests to compare the accent conversions using the equivalent articulatory synthesis of the acoustic-based method against the articulatory-based accent conversion. For this comparative study, we used the corpus described in section 5.2, which contains audio and articulatory recordings from a native speaker and a non-native speaker. As described in that same section, a set of 294 sentences were used for training and the remaining 50 sentences for testing purpose. Similarly, STRAIGHT was used to extract acoustic features. After feature extraction, the acoustic feature vectors consisted of $MFCC_{1-24}$; and the articulatory feature vectors consisted of six EMA positions, $MFCC_0$, *nasality* and $log(f_0)$.

To match the number of mixture components of the GMMs with that of the articulatory-based method, we also trained the cross speaker forward mappings with 128 mixture components.

7.3.1 Experimental conditions

We considered four experimental conditions for the listening tests: (i) the proposed *equivalent articulatory synthesis* of acoustic-based accent conversion (AC_{EQV}), (ii) the articulatory-based accent conversion using the method described in section 5.1 (AC_{EMA}), (iii) MFCC compression of L2 speech ($L2_{MFCC}$), and (iv) guise of L1 utterances to match the vocal tract length and the pitch range of L2 ($L1_{GUISE}$). See section 5.2.1 for more details on the last three conditions (AC_{EMA}, $L2_{MFCC}$ and $L1_{GUISE}$).

7.4 Results

We performed three listening experiments to compare AC_{EQV} and AC_{EMA} in terms of the perceived reduction in non-native accents, intelligibility, and the voice-similarity with the L2 speaker. In the first experiment, we performed a forced pairwise comparison test to identify the most native-like accent conversion. In the second experiment, we evaluated the intelligibility of AC_{EQV}, and compared against the intelligibility of AC_{EMA}. In the third and final experiment, we compared if the AC_{EQV} preserves the voice-similarity of the L2 speaker.

As before, the participants for all the listening tests were recruited through Mechanical Turk, Amazon's online crowdsourcing tool; —see section 5.2.2 for more detail.

7.4.1 Non-native accent evaluation

In a first listening experiment we sought to compare the perceived reduction of non-native accents between the two foreign accent conversion strategies. For this purpose, participants were asked to listen to a pair of utterances of the same sentence from AC_{EQV} and AC_{EMA}, and select the most native-like among them. We tested on the same subset of 15 test sentences in section 5.3.3 so that the results could be compared.

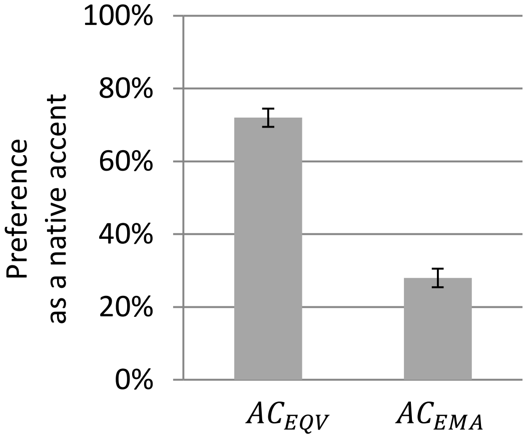


Figure 33: Subjective evaluation of accentedness. Participants selected the most native-like utterances between AC_{EQV} vs. AC_{EMA}.

Participants listened to 30 pairs of utterances (15 $AC_{EMA} - AC_{EQV}$ pairs and 15 $AC_{EQV} - AC_{EMA}$ pairs) presented in random order to account for ordering effects. As

Figure 33 shows, the participants rated AC_{EQV} more native-like than AC_{EMA} in 72% $(s.e = 3\%)$ of the sentences, which is significantly higher $(t(14) = 8.87, p < 0.001)$ than the 50% chance level. This result shows that the acoustic-based strategy is more effective than articulatory-based strategy in reducing non-native accents.

7.4.2 Intelligibility assessment

In a second experiment, we assessed the intelligibility of AC_{EQV} to compare against the similar assessment of AC_{EMA} in section 5.3.2. Following the same approach described in section 5.3.2, a group of native speakers of American English (N=15 each) were asked to transcribe the 46 test utterances[19] from the experimental condition AC_{EQV}. From the transcription, we calculated word accuracy (W_{acc}) as the ratio of the number of correctly identified words to the total number of words in the utterance. Participants also rated the (subjective) intelligibility of the utterances (S_{intel}) using a 7-point Likert scale (1: not intelligible at all, 3: somewhat intelligible, 5: quite a bit intelligible, and 7: extremely intelligible).

[19] Four of 50 test sentences for the L2 speaker had missing EMA data and were removed from the analysis.

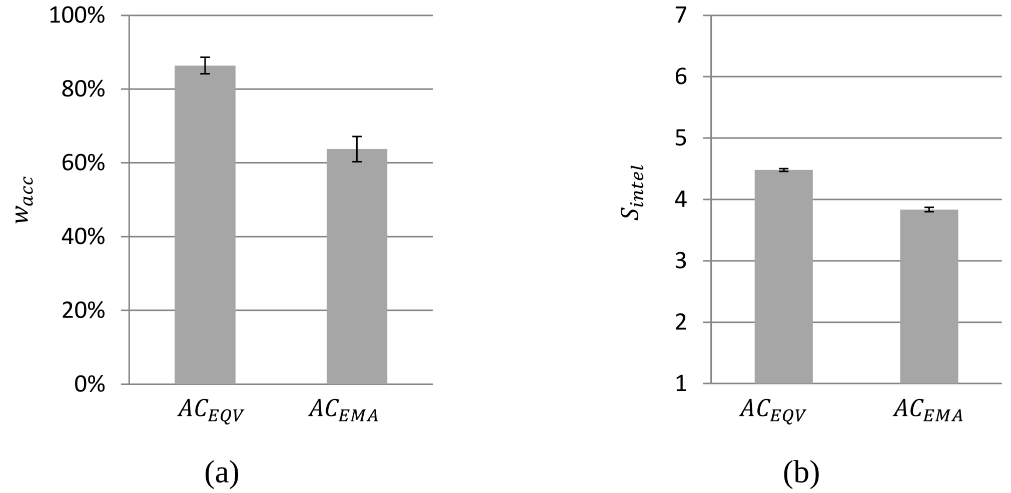


Figure 34: (a) Word accuracy and (b) subjective intelligibility ratings for AC_{EQV} and AC_{EMA}.

Figure 34 shows the word accuracy and intelligibility ratings for AC_{EQV} against that of the articulatory-based accent conversion (AC_{EMA}) from section 5.3.2. The results show that the accent conversions in the acoustic domain $\left(AC_{EQV}: W_{acc} = 0.86,\ S_{intel} = 4.48\right)$ were rated significantly more intelligible $(p < 0.001; t - test)$ than the conversion in the articulatory domain $(AC_{EMA}: W_{acc} = 0.64,\ S_{intel} = 3.83)$. Since accent conversions in both the groups are driven by the same articulatory input features, the higher intelligibility ratings for AC_{EQV} than AC_{EMA} may be due to higher reduction in the perceived non-native accentedness in AC_{EQV}.

7.4.3 Voice identity assessment

In a third and final listening experiment, we evaluated if the articulatory equivalent synthesis of acoustic-based foreign accent conversion was able to preserve the voice identity of the L2 speaker. For this purpose, participants were asked to compare the voice similarity between pairs of utterances, one from AC_{EQV} , the other from $L2_{MFCC}$. As a sanity check we also included the pairs of utterances from $L2_{MFCC}$ and $L1_{GUISE}$, the latter being a simple guise of L1 utterances that matches the pitch range and vocal tract length of the L2 speaker. As in the prior voice-similarity tests, the two sentences on each pair were linguistically different, and the presentation order was randomized for conditions within each pair and for pairs of conditions. Participants ($N = 15$) rated 40 pairs, 20 from each group ($L2_{MFCC} - AC_{EQV}$, $L2_{MFCC} - L1_{GUISE}$) randomly interleaved, and were asked to (i) determine if the utterances were from the same or a different speaker and (ii) rate how confident they were in their assessment using a seven-point Likert scale (1: not confident at all, 3: somewhat confident, 5: quite a bit confident, and 7: extremely confident). The responses and their confidence ratings were then combined to form a voice similarity score (VSS) ranging from -7 (extremely confident they are different speaker) to $+7$ (extremely confident they are from the same speaker).

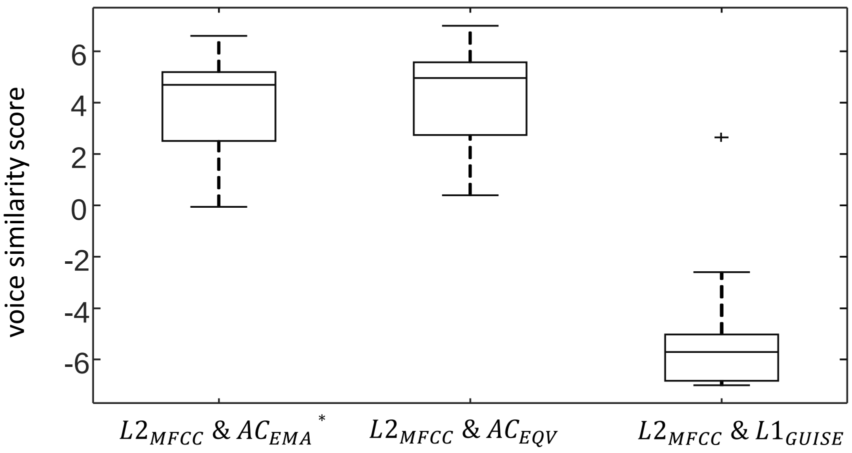


Figure 35: Average pairwise voice similarity scores. $L2_{MFCC}$& AC_{EMA} * from section 6.5.3.

Figure 35 shows the boxplot of average VSS between pairs of experimental conditions. Participants were 'quite' confident ($VSS = 4.2, s.e. = 0.5$) that the $L2_{MFCC}$ and AC_{EQV} were from the same speaker, suggesting that the equivalent articulatory synthesis for the acoustic-based strategy method successfully preserved the voice-identity of L2 speaker. The VSS was also found comparable ($t(28) = 0.32, p = 0.74, two - tail$) to the VSS between AC_{EMA} and $L2_{MFCC}$ ($VSS = 4.0, s.e. = 0.5$) reported for the articulatory-based method in section 5.3.4. Moreover, the participants were also 'quite' confident that ($VSS = -5.06, s.e. = 0.7$) the $L2_{MFCC}$ and $L1_{GUISE}$ were from different speakers, corroborating our prior finding (section 5.3.4) that a simple guise of L1 utterances is not sufficient to match the voice of the L2 speaker.

7.5 Discussions

In this section we compared two foreign accent conversion strategies based on acoustic-based and articulatory-based modifications. Since the articulatory-based method is driven by articulatory features, which is a partial representation of the vocal tract and the less informative of the phonetic variability than the acoustic features, the direct comparison between the two methods in terms of perceived non-native accents can be biased. To avoid such issue, we built an equivalent articulatory synthesizer for the acoustic-based method, so that both methods use articulatory features from a reference native speaker as the carrier of the native linguistic gestures. Perceptual listening tests indicate that the acoustic-based strategy (the equivalent articulatory synthesis) is more effective in reducing perceived non-native accents than the articulatory-based strategy. The acoustic-based accent conversion was also found more intelligible than the articulatory-based conversion. These findings make the acoustic-based methods even more appealing for computer aided pronunciation tool than the expensive articulatory-based methods.

After accounting for the differences in the representation of the input linguistic gestures, the two strategies differed only in the way accent-related differences between L1 and L2 are addressed. In the articulatory-based strategy, accent modification is performed in articulatory domain only using Procrustes transforms of EMA pellet positions to account for differences in the vocal tract geometry of the two speakers. On the other hand, in the acoustic-based strategy a GMM-based mapping of the acoustic feature spaces is used. Our finding suggests that the accent modification is more effective in acoustic space, but further study is required to verify if the comparatively lower reduction in perceived non-native accents is due to the partial representation of vocal tract. Even after the inclusion of voicing and nasality features, the EMA data does not have the same level of phonetic information as the acoustic features. Having articulatory representations such as rt-MRI (Narayanan *et al.*, 2011), which contains the 3D image of the complete vocal tract, may improve the performance of articulatory-based strategies.

8. CONCLUSIONS

8.1 Summary

We developed statistical parametric techniques to generate speech with native accent but the voice of a non-native speaker. The techniques were developed for both the acoustic and articulatory domains. In the proposed acoustic-based method, we estimate the equivalent L2 acoustic features from the acoustic features of a reference native utterance using a cross-speaker spectral mapping. The GMM-based mappings were trained on the joint distribution of L1 and L2 acoustic feature vectors paired with each other based on acoustic similarity, unlike the force-aligned pairs used in conventional voice conversion. The results show a perceivable reduction in non-native accents. Most importantly, the method was also able to preserve the voice-identity of the non-native speaker unlike existing vocoding approaches (Felps *et al.*, 2009; Aryal and Gutierrez-Osuna, 2013). In the articulatory-based methods, we used articulatory data from a native reference utterance to drive the statistical parametric articulatory synthesizer build for a non-native speaker. Unlike unit-selection used in the only existing articulatory-based approach (Felps *et al.*, 2012), the statistical parametric synthesizer has lower data requirement and enough flexibility to generate novel sounds. We evaluated two statistical parametric synthesis models for articulatory-based accent conversion. First, we used GMM-based articulatory synthesizer because of their proven flexibility to interpolate novel sounds, and found it effective in reducing the non-native accents. However, the GMM-based method uses an expensive trajectory optimization stage, which considers the dynamics of acoustic features to reduce spectral discontinuities. Therefore, we proposed a new articulatory synthesis model based on deep neural networks (DNN). The DNN-based synthesizer exploits the temporal nature of speech using the contextualized articulatory features as the input and obviates the need for the expensive trajectory optimization of the estimated acoustic features. The run-time of a Matlab implementation of the DNN-based synthesizer in a typical modern-day personal computer was found to be lower than the frame-rate making the method suitable for real-time conversion. From listening tests, we also found that the DNN-based method had higher reduction of perceived non-native accents and superior acoustic than the GMM-based method.

Given the high expense and the difficulty of collecting articulatory data, we compared the articulatory-based strategy against a more practical acoustic-based strategy. Because of their differences in acoustic quality known to affect the perceptual evaluation of non-native accentedness, we built an equivalent articulatory synthesizer for the acoustic-based accent conversion method. In a listening test comparing the GMM-based articulatory accent conversion against the output of the equivalent articulatory synthesizer of the acoustic-based strategy, we found the acoustic-based strategy more effective in reducing the perceived non-native accent. Given the lower cost of the acoustic-based accent conversion method, these finding make them even more appealing for the computer aided pronunciation training tools. However, further study is required to investigate if the non-native accent-reduction in the articulatory-based strategy can surpass the performance of the acoustic-based method, if a complete representation of the vocal-tract anatomy such as rt-MRI (Narayanan *et al.*, 2011) were used, instead of the partial representation used in this study.

8.2 Main contributions

The main contributions of this dissertation work are as follows:

- Development of an acoustic-based foreign accent conversion method immune to the ‘third-speaker’ problem and the difficulties in aligning native and non-native speech —the main limitations of the existing vocoding-based acoustic methods.
- Creation of an articulatory technique to transpose accents from a reference L1 speaker to an L2 speaker by driving a statistical parametric articulatory synthesizer for the L2 speaker with the articulatory gestures from the L1 speaker.
- Development of a DNN-based articulatory parametric synthesizer suitable for real-time accent conversion.
- Demonstration that the exploitation of temporal nature of speech in contextualized articulatory input via deep neural networks is more computationally efficient than using trajectory optimization of estimated acoustic features in GMM-based synthesis. We also demonstrated that the efficiency comes without compromising the model’s ability to reduce perceived non-native accents.
- Designed a method to compare the acoustic-based strategy against the less practical but theoretically sound articulatory-based strategy in terms of their ability to reduce perceived non-native accents, accounting for their differences in acoustic quality known to impact the accent perception.

- Demonstrated that the acoustic-based strategy is more effective in reducing non-native accents than the articulatory-based strategy.

8.3 Future Work

8.3.1 Large scale validation

Due to the rarity of parallel articulatory recordings from L1 and L2 speakers, we validated our methods in a single L2 speaker. As the articulatory recordings from L2 speakers becomes more accessible, these methods need to be validated for multiple L2 speakers with different native languages, speaking styles and levels of proficiency. An interesting new resource in this regard is the Marquette University Electromagnetic Articulography Mandarin Accented English (EMA-MAE), which contains a large EMA corpus from multiple Mandarin second-language speakers of American English (Ji *et al.*, 2014). This new resource makes it possible to validate our articulatory synthesis and accent conversion methods across multiple speakers.

8.3.2 Performance improvement

There are several ways we can improve the performance of the foreign accent conversion methods described in this work. In the case of our articulatory-based methods, the synthesis quality was affected due to the partial representation of vocal tract via EMA position data. Future work may extend this study using the more informative articulatory representation provided by real-time magnetic resonance imaging (rt-MRI) (Narayanan *et al.*, 2011). In comparison to EMA, which only captures a few fleshpoints in the frontal oral cavity, rt-MRI provides information about the entire vocal tract, from lips to glottis, which may result in more intelligible and native-like accent conversions. Similarly, collecting supplementary data on tongue palate closure via electropalatography may also improve intelligibility, especially in the case of stops. Another possible approach is to transform the articulatory measurements into constriction-based representation such as TVs because they are known to have less variability than EMA pellet positions(McGowan, 1994; Mitra *et al.*, 2011).

Phonetic information can be used to further improve the performance of foreign accent conversion. The phonetic information adds prior to the cross-speaker articulatory mappings and the articulatory-to-acoustic mappings, the main building blocks of the articulatory-based methods. For example, cross-speaker articulatory mappings may cause a lingua-alveolar stop to become fricative due to a small error in the tongue-tip height. Such errors can be minimized if the mappings are aware of the characteristics of the target phone. In the case of forward mappings, Felps *et al.* (2010) have shown that the accuracy of articulatory-to-acoustic mappings can be increased by using phone-specific weights for the EMA pellet positions of critical articulators. Furthermore, language specific knowledge can help optimize both the mappings to reduce errors that have high functional load; e.g., contrast between initial /p/ and /b/ has relatively higher functional load compared to contrast between final /t/ and /d/ (Jesse, 2012).

The naturalness of the accent conversions in this work is also affected by the smoothing effects inherent to statistical mappings. The estimation model incorporates global variances to reduce the smoothing effect but future work may explore the use of modulation spectrum-based post filtering which is known to improve naturalness in voice conversion (Takamichi *et al.*, 2014). Similarly, the exemplar-based voice conversion technique (Takashima *et al.*, 2012) known for their close to human-like acoustic quality can be adapted for accent conversion.

At present, our approach uses L1 aperiodicity spectra and therefore does not consider speaker individuality cues that may be contained in the L2 aperiodicity (Kawahara, 1997). Thus, further improvements in voice similarity may be obtained by replacing the L1 aperiodicity with its L2 equivalent. One possibility is to estimate L2 aperiodicity from the estimated L2 spectra by exploiting the relation between both signals (Silén *et al.*, 2011).

8.3.3 Application of foreign accent conversion methods in computer aided pronunciation training

The main motivation behind the development of foreign accent conversion methods is their application in computer-aided pronunciation training for non-native learners. Studies have shown that the pronunciation training is more effective when the teacher's voice matches the learner's (Nagano and Ozawa, 1990; Probst *et al.*, 2002; Bissiri *et al.*, 2006). Due to the ease of modifying prosody, the effect of using the learner's own voice (instead of finding a teacher with similar voice) in training prosody has been well studied (Nagano and Ozawa, 1990), leading towards the development of automatic prosody modification techniques for the computer assisted pronunciation training tools (Sundström, 1998). However, the effect of using the learner's own voice in training segmental aspects (e.g. vowel quality) of accents has not been studied. With the foreign accent conversion methods developed in this work, it is possible to evaluate the effect of using the learner own utterances following the reduction of non-native accents in training the segmental aspects of accent.

8.3.4 Extension to other articulatory speech modification problems

Using articulatory speech synthesis has been suggested as one of the promising techniques for expressive synthesis (Lee *et al.*, 2005; Schröder, 2009). Schröder (2009) suggests using the physical synthesis model of Birkholz (2007) that is capable of generating intelligible speech from a ‘score’ representation of articulatory movement. As an alternative, we propose adapting the articulatory techniques developed in this work to modify emotions. Our data-driven approach is more practical than the physical model if the expressive synthesizer is built for a specific speaker. Similarly, our articulatory techniques can also be extended to generate signing voices (Birkholz, 2007).

REFERENCES

Abe, M., Nakamura, S., Shikano, K., and Kuwabara, H. (**1988**). "Voice conversion through vector quantization," in *Proceedings of ICASSP*, pp. 655-658.

Abercrombie, D. (**1949**). "Teaching pronunciation," ELT Journal **3**, 113-122.

Abrahamsson, N., and Hyltenstam, K. (**2008**). "The robustness of aptitude effects in near-native second language acquisition," Studies in second language acquisition **30**, 481-509.

Al Bawab, Z., Bhiksha, R., and Stern, R. M. (**2008**). "Analysis-by-synthesis features for speech recognition," in *Proceedings of ICASSP*, pp. 4185-4188.

Andrew, G., Arora, R., Bilmes, J., and Livescu, K. (**2013**). "Deep canonical correlation analysis," in *Proceedings of ICML*, pp. 1247-1255.

Arora, R., and Livescu, K. (**2013**). "Multi-view CCA-based acoustic features for phonetic recognition across speakers and domains," in *Proceedings of ICASSP*, pp. 7135-7139.

Arslan, L. M., and Talkin, D. (**1997**). "Voice Conversion By Codebook Mapping Of Line Spectral Frequencies And Excitation Spectrum," in *Proceedings of EUROSPEECH*, pp. 1347–1350.

Aryal, S., Felps, D., and Gutierrez-Osuna, R. (**2013**). "Foreign accent conversion through voice morphing," in *Proceedings of INTERSPEECH*, pp. 3077-3081.

Aryal, S., and Gutierrez-Osuna, R. (**2013**). "Articulatory inversion and synthesis: towards articulatory-based modification of speech," in *Proceedings of ICASSP*, pp. 7952-7956.

Aryal, S., and Gutierrez-Osuna, R. (**2014a**). "Accent conversion through cross-speaker articulatory synthesis," in *Proceedings of ICASSP*, pp. 7744-7748.

Aryal, S., and Gutierrez-Osuna, R. (**2014b**). "Can voice conversion be used to reduce non-native accents?," in *Proceedings of ICASSP*, pp. 7929-7933.

Aryal, S., and Gutierrez-Osuna, R. (**2015a**). "Articulatory-based conversion of foreign accents with deep neural networks," in *Proceedings of INTERSPEECH*, pp. 3385-3389.

Aryal, S., and Gutierrez-Osuna, R. (**2015b**). "Reduction of non-native accents through statistical parametric articulatory synthesis," J. Acoust. Soc. Am. **137**, 433-446.

Aryal, S., and Gutierrez-Osuna, R. (**2015 (in press)**). "Data driven articulatory synthesis with deep neural networks," Computer Speech & Language.

Atal, B. S., Chang, J. J., Mathews, M. V., and Tukey, J. W. (**1978**). "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," J Acoust. Soc. Am. **63**, 1535-1555.

Atal, B. S., and Hanauer, S. L. (**1971**). "Speech analysis and synthesis by linear prediction of the speech wave," J Acoust. Soc. Am. **50**, 637-655.

Bao, Y., Jiang, H., Liu, C., Hu, Y., and Dai, L. (**2012**). "Investigation on dimensionality reduction of concatenated features with deep neural network for LVCSR systems," in *Proceedings of ICSP*, pp. 562-566.

Barefoot, S. M., Bochner, J. H., Johnson, B. A., and vom Eigen, B. A. (**1993**). "Rating Deaf Speakers' ComprehensibilityAn Exploratory Investigation," American Journal of Speech-Language Pathology **2**, 31-35.

Birkholz, P. (**2007**). "Articulatory synthesis of singing," in *Proceedings of INTERSPEECH*, pp. 4001-4004.

Birkholz, P., and Jackel, D. (**2003**). "A three-dimensional model of the vocal tract for speech synthesis," in *Proceedings of the ICPhS*, pp. 2597-2600.

Birkholz, P., Jackèl, D., and Kroger, B. (**2006**). "Construction and control of a three-dimensional vocal tract model," in *Proceedings of ICASSP*, pp. 873–876.

Bissiri, M. P., Pfitzinger, H. R., and Tillmann, H. G. (**2006**). "Lexical stress training of German compounds for Italian speakers by means of resynthesis and emphasis," in *Proceedings of AICSST*, pp. 24-29.

Black, J. W. (**1973**). "The “phonemic” content of backward-reproduced speech," J. of Speech, Lang., and Hearing Research **16**, 165-174.

Brennan, E. M., and Brennan, J. S. (**1981**). "Measurements of accent and attitude toward Mexican-American speech," Journal of Psycholinguistic Research **10**, 487-501.

Browman, C. P., and Goldstein, L. (**1990**). "Gestural specification using dynamically-defined articulatory structures," J. Phonetics **18**, 299-320.

Browman, C. P., and Goldstein, L. (**1992**). "Articulatory phonology: An overview," Phonetica **49**, 155-180.

Campbell-Kibler, K. (**2009**). "The nature of sociolinguistic perception," Language Variation and Change **21**, 135-156.

Campbell, N. (**1998**). "Foreign-language speech synthesis," in *Proceedings SSW3*, pp. 177-180.

Celce-Murcia, M., Brinton, D. M., and Goodwin, J. M. (**1996**). *Teaching pronunciation: A reference for teachers of English to speakers of other languages* (Cambridge University Press).

Cho, K. H. (**2013**). "Matlab code for restricted/deep Boltzmann machines and autoencoders," (https://github.com/kyunghyuncho/deepmat).

Cho, K. H., Raiko, T., and Ilin, A. (**2013**). "Gaussian-Bernoulli deep Boltzmann machine," in *Proceedings of IJCNN*, pp. 1-7.

Crawford, W. (**1987**). "The pronunciation monitor: L2 acquisition considerations and pedagogical priorities," in *Current perspectives on pronunciation: Practices anchored in theory.* , edited by J. Morley (TESOL), pp. 101-121.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. (**2005**). "Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences," Journal of Experimental Psychology: General **134**, 222-241.

Davis, S. B., and Mermelstein, P. (**1980**). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Audio Speech Lang. Process. **28**, 357-366.

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., and Brumberg, J. S. (**2010**). "Silent speech interfaces," Speech Commun. **52**, 270-287.

Derwing, T. M., Munro, M. J., and Wiebe, G. (**1998**). "Evidence in favor of a broad framework for pronunciation instruction," Language Learning **48**, 393-410.

Desai, S., Black, A. W., Yegnanarayana, B., and Prahallad, K. (**2010**). "Spectral mapping using artificial neural networks for voice conversion," IEEE Trans. Audio Speech Lang. Process. **18**, 954-964.

Dovidio, J. F., and Fiske, S. T. (**2012**). "Under the Radar: How Unexamined Biases in Decision-Making Processes in Clinical Interactions Can Contribute to Health Care Disparities," American Journal of Public Health **102**, 945-952.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., and Bengio, S. (**2010**). "Why does unsupervised pre-training help deep learning?," J. Mach. Learn. Res. **11**, 625-660.

Eskenazi, M. (**1999**). "Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype," Language learning & technology **2**, 62-76.

Fant, G. (**1970**). *Acoustic Theory of Speech Production* (Mouton, The Hague).

Fayer, J. M., and Krasinski, E. (**1987**). "Native and nonnative judgments of intelligibility and irritation," Language Learning **37**, 313-326.

Felps, D. (**2011**). "ARTICULATORY-BASED SPEECH PROCESSING METHODS FOR FOREIGN ACCENT CONVERSION," (Texas A&M University).

Felps, D., Aryal, S., and Gutierrez-Osuna, R. (**2014**). "Normalization of articulatory data through Procrustes transformations and analysis-by-synthesis," in *Proceedings of ICASSP*, pp. 3051-3055.

Felps, D., Bortfeld, H., and Gutierrez-Osuna, R. (**2009**). "Foreign accent conversion in computer assisted pronunciation training," Speech commun. **51**, 920-932.

Felps, D., Geng, C., Berger, M., Richmond, K., and Gutierrez-Osuna, R. (**2010**). "Relying on critical articulators to estimate vocal tract spectra in an articulatory-acoustic database," in *Proceedings of INTERSPEECH*, pp. 1990-1993.

Felps, D., Geng, C., and Gutierrez-Osuna, R. (**2012**). "Foreign accent conversion through concatenative synthesis in the articulatory domain," IEEE Trans. Audio Speech Lang. Process. **20**, 2301-2312.

Felps, D., and Gutierrez-Osuna, R. (**2010**). "Developing objective measures of foreign-accent conversion," IEEE Trans. Audio Speech Lang. Process. **18**, 1030-1040.

Geng, C., and Mooshammer, C. (**2009**). "How to stretch and shrink vowel systems: Results from a vowel normalization procedure," J. Acoust. Soc. Am. **125**, 3278-3288.

Ghosh, P., and Narayanan, S. (**2011a**). "Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion," J Acoust. Soc. Am. **130**, EL251-EL257.

Ghosh, P. K., and Narayanan, S. S. (**2011b**). "A subject-independent acoustic-to-articulatory inversion," in *Proceedings of ICASSP*, pp. 4624-4627.

Giles, H. (**1970**). "Evaluative reactions to accents," Educational review **22**, 211-227.

Gluszek, A., and Dovidio, J. F. (**2010**). "The way they speak: a social psychological perspective on the stigma of non-native accents in communication," Personality and Social Psychology Review **14**, 214-237.

Goldstein, B. (**2001**). "Transcription of Spanish and Spanish-influenced English," Communication Disorders Quarterly **23**, 54-60.

Goldstein, M. (**1995**). "Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener," Speech commun. **16**, 225-244.

Guenther, F. H. (**1994**). "A neural network model of speech acquisition and motor equivalent speech production," Biological cybernetics **72**, 43-53.

Hansen, J. H., and Arslan, L. M. (**1995**). "Foreign accent classification using source generator based prosodic features," in *Proceedings of ICASSP*, pp. 836-839.

Hashi, M., Westbury, J. R., and Honda, K. (**1998**). "Vowel posture normalization," J Acoust. Soc. Am. **104**, 2426-2437.

Helman, L. A. (**2004**). "Building on the sound system of Spanish: Insights from the alphabetic spellings of English-language learners," The Reading Teacher, 452-460.

Hermansky, H., and Broad, D. J. (**1989**). "The effective second formant F2' and the vocal tract front-cavity," in *Proceedings of ICASSP*, pp. 480-483.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., and Sainath, T. N. (**2012**). "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Processing Magazine **29**, 82-97.

Hinton, G., and Sejnowski, T. (**1983**). "Optimal perceptual inference," in *Proceedings of ICCVPR*, pp. 448-453.

Hinton, G. E. (**2012**). "A practical guide to training restricted Boltzmann machines," in *Neural Networks: Tricks of the Trade* (Springer), pp. 599-619.

Hiroya, S., and Honda, M. (**2004**). "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," IEEE Trans. Speech Audio Process. **12**, 175-185.

Huckvale, M. (**2004**). "ACCDIST: a metric for comparing speakers' accents," in *Proceedings of INTERSPEECH*, pp. 29-32.

Huckvale, M., and Yanagisawa, K. (**2007**). "Spoken language conversion with accent morphing," in *Proceedings of ISCA SSW*, pp. 64-70.

Hunt, A. J., and Black, A. W. (**1996**). "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP*, pp. 373-376.

Imai, S. (**1983**). "Cepstral analysis synthesis on the mel frequency scale," in *Proceedings of ICASSP*, pp. 93--96.

ITU-T (**2003**). "Recommendation G.114: One-way transmission time."

ITU-T (**2004**). "Recommendation P.563: Single-ended method for objective speech quality assessment in narrow-band telephony applications."

ITU-T (**2006**). "Recommendation P.800: Methods for subjective determination of transmission quality."

Jenner, B. R. A. (**1976**). "Interlanguage and Foreign Accent," Interlanguage Studies Bulletin **1**, 166-195.

Jesse, G. (**2012**). "Beaches and peaches: Common pronunciation errors among L1 Spanish speakers of English," in *Proceedings of PSLLT*, pp. 205-215.

Ji, A., Berry, J., and Johnson, M. T. (**2014**). "The Electromagnetic Articulography Mandarin Accented English (EMA-MAE) Corpus of Acoustic and 3D Articulatory Kinematic Data," in *Proceedings of ICASSP*, pp. 7769-7773.

Kaburagi, T., and Honda, M. (**1998**). "Determination of the vocal tract spectrum from the articulatory movements based on the search of an articulatory-acoustic database," in *Proceedings of ICSLP*, pp. 433-436.

Kain, A., and Macon, M. W. (**1998**). "Spectral voice conversion for text-to-speech synthesis," in *Proceedings of ICASSP*, pp. 285-288.

Kain, A., and Macon, M. W. (**2001**). "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," in *Proceedings of ICASSP*, pp. 813-816.

Kalin, R., and Rayko, D. S. (**1978**). "Discrimination in evaluative judgments against foreign-accented job candidates," Psychological Reports **43**, 1203-1209.

Kawahara, H. (**1997**). "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *Proceedings of ICASSP*, pp. 1303-1306.

Kello, C. T., and Plaut, D. C. (**2004**). "A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters," J Acoust. Soc. Am. **116**, 2354-2364.

Kempen, G. (**1992**). "Second language acquisition as a hybrid learning process," in *Cognitive modelling and interactive environments in language learning* (Springer), pp. 139-144.

King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., and Wester, M. (**2007**). "Speech production knowledge in automatic speech recognition," J Acoust. Soc. Am. **121**, 723-742.

Kreiman, J., and Papcun, G. (**1991**). "Comparing discrimination and recognition of unfamiliar voices," Speech Commun. **10**, 265-275.

Kruskal, J. B. (**1964**). "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," Psychometrika **29**, 1-27.

Ladefoged, P. (**1980**). "Articulatory parameters," Language and Speech **23**, 25-30.

Lane, H. (**1963**). "Foreign accent and speech distortion," J. Acoust. Soc. Am. **35**, 451-453.

Lavner, Y., Gath, I., and Rosenhouse, J. (**2000**). "The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels," Speech Commun. **30**, 9-26.

Lee, S., Yildirim, S., Kazemzadeh, A., and Narayanan, S. (**2005**). "An articulatory study of emotional speech production," in *Proceedings of INTERSPEECH*, pp. 497-500.

Lenneberg, E. H. (**1967**). *Biological foundations of language* (Wiley, New York).

Livescu, K., Cetin, O., Hasegawa-Johnson, M., King, S., Bartels, C., Borges, N., Kantor, A., Lal, P., Yung, L., Bezman, A., and others (**2007**). "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU summer workshop," in *Prodeedings of ICASSP*, pp. 621-624.

Livescu, K., and Stoehr, M. (**2009**). "Multi-view learning of acoustic features for speaker recognition," in *Proceedings of ASRU*, pp. 82-86.

Maeda, S. (**1979**). "An articulatory model of the tongue based on a statistical analysis," J Acoust. Soc. Am. **65**, S22.

Maeda, S. (**1982**). "A digital simulation method of the vocal-tract system," Speech commun. **1**, 199-229.

Maeda, S. (**1990**). "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model," in *Speech production and speech modelling*, edited by W. Hardcastle, and A. Marchal (Amsterdam: Kluwer Academic Publisher), pp. 131-149.

Maragakis, M. G., and Potamianos, A. (**2008**). "Region-based vocal tract length normalization for ASR," in *Proceedings of INTERSPEECH*, pp. 1365-1368.

McAulay, R., and Quatieri, T. (**1986**). "Speech analysis/synthesis based on a sinusoidal representation," IEEE Trans. Audio Speech Lang. Process. **34**, 744--754.

McGowan, R. S. (**1994**). "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests," Speech Commun. **14**, 19-48.

Mermelstein, P. (**1973**). "Articulatory model for the study of speech production," J. Acoust. Soc. Am. **53**, 1070-1082.

Mitra, V., Nam, H., Espy-Wilson, C., Saltzman, E., and Goldstein, L. (**2010**). "Robust word recognition using articulatory trajectories and Gestures," in *Proceedings of INTERSPEECH*, pp. 2038-2041.

Mitra, V., Nam, H., Espy-Wilson, C. Y., Saltzman, E., and Goldstein, L. (**2011**). "Speech inversion: Benefits of tract variables over pellet trajectories," in *Proceedings of ICASSP*, pp. 5188-5191.

Morley, J. (**1991**). "The pronunciation component in teaching English to speakers of other languages," Tesol Quarterly, 481-520.

Moulines, E., and Charpentier, F. (**1990**). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Communication **9**, 453-467.

Munro, M., and Derwing, T. (**1995**). "Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners," Language Learning & Technology **45**, 73-97.

Munro, M. J., Derwing, T. M., and Burgess, C. S. (**2010**). "Detection of nonnative speaker status from content-masked speech," Speech Commun. **52**, 626-637.

Muramatsu, T., Ohtani, Y., Toda, T., Saruwatari, H., and Shikano, K. (**2008**). "Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory," in *Proceedings of INTERSPEECH*, pp. 1076-1079.

Nabney, I. T. (**2002**). *NETLAB: Algorithms for Pattern Recognition* (Springer).

Nagano, K., and Ozawa, K. (**1990**). "English speech training using voice conversion," in *Proceedings of ICSLP* (Kobe, Japan), pp. 1169-1172.

Nakamura, K., Toda, T., Nankaku, Y., and Tokuda, K. (**2006**). "On the use of phonetic information for mapping from articulatory movements to vocal tract spectrum," in *Proceedings of ICASSP*, pp. I93-I96.

Narayanan, S., Bresch, E., Ghosh, P. K., Goldstein, L., Katsamanis, A., Kim, Y., Lammert, A. C., Proctor, M. I., Ramanarayanan, V., and Zhu, Y. (**2011**). "A Multimodal Real-Time MRI Articulatory Corpus for Speech Research," in *Proceedings of INTERSPEECH*, pp. 837-840.

Neufeld, G. (**1978**). "On the Acquisition of Prosodic and Articulatory Features in Adult Language Learning," Canadian Modern Language Review **34**, 163-174.

Ohtani, Y., Toda, T., Saruwatari, H., and Shikano, K. (**2006**). "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," in *Proceedings of ICSLP*, pp. 2266-2269.

Panchapagesan, S., and Alwan, A. (**2009**). "Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC," Computer speech & language **23**, 42-64.

Paul, D. (**1981**). "The spectral envelope estimation vocoder," IEEE Trans. Audio Speech Lang. Process. **29**, 786-794.

Peabody, M. A. (**2011**). "Methods for Pronunciation Assessment in Computer Aided Language Learning."

Pisoni, D. B., and Hunnicutt, S. (**1980**). "Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system," in *Proceedings of ICASSP*, pp. 572-575.

Probst, K., Ke, Y., and Eskenazi, M. (**2002**). "Enhancing foreign language tutors–in search of the golden speaker," Speech Commun. **37**, 161-173.

Pruthi, T., and Espy-Wilson, C. Y. (**2004**). "Acoustic parameters for automatic detection of nasal manner," Speech Commun. **43**, 225-239.

Qin, C., and Carreira-Perpinán, M. A. (**2007a**). "A comparison of acoustic features for articulatory inversion," in *Interspeech*, pp. 2469-2472.

Qin, C., and Carreira-Perpinán, M. A. (**2007b**). "An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping," in *Proceedings of INTERSPEECH*, pp. 2300-2303.

Qin, C., and Carreira-Perpinán, M. A. (**2009**). "Adaptation of a predictive model of tongue shapes," in *Proceedings of INTERSPEECH*, pp. 772-775.

Qin, C., Carreira-Perpinán, M. A., Richmond, K., Wrench, A., and Renals, S. (**2008**). "Predicting tongue shapes from a few landmark locations," in *Proceedings of INTERSPEECH* (Brisbane, Australia), pp. 2306-2309.

Repp, B. H., and Williams, D. R. (**1987**). "Categorical tendencies in imitating self-produced isolated vowels," Speech Commun. **6**, 1-14.

Richmond, K., King, S., and Taylor, P. (**2003**). "Modelling the uncertainty in recovering articulation from acoustics," Computer Speech & Language **17**, 153-172.

Rogers, C. L., and Dalby, J. M. (**1996**). "Prediction of foreign-accented speech intelligibility from segmental contrast measures," J. Acoust. Soc. Am. **100**, 2725-2725.

Rubin, D. L., and Smith, K. A. (**1990**). "Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of nonnative English-speaking teaching assistants," International Journal of Intercultural Relations **14**, 337-353.

Rubin, P., Baer, T., and Mermelstein, P. (**1981**). "An articulatory synthesizer for perceptual research," J Acoust. Soc. Am. **70**, 321-328.

Rumelhart, D., Hinton, G., and Williams, R. (**1986**). "Learning representations by back-propagating errors," Nature **323**, 533-536.

Ryan, E. B., and Carranza, M. A. (**1975**). "Evaluative reactions of adolescents toward speakers of standard English and Mexican American accented English," Journal of personality and social psychology **31**, 855.

Salakhutdinov, R., and Hinton, G. E. (**2009**). "Deep Boltzmann machines," in *Proceedings of AISTATS*, pp. 448-455.

Saltzman, E. L., and Munhall, K. G. (**1989**). "A dynamical approach to gestural patterning in speech production," Ecological psychology **1**, 333-382.

Sato, C. (**1991**). "Sociolinguistic variation and language attitudes in Hawaii," English around the world: Sociolinguistic perspectives, 647-663.

Schröder, M. (**2009**). "Expressive speech synthesis: Past, present, and possible futures," in *Affective information processing* (Springer), pp. 111-126.

Scovel, T. (**1969**). "Foreign accents, language acquisition, and cerebral dominance," Language Learning **19**, 245-253.

Scovel, T. (**1988**). *A Time to Speak: A Psycholinguistic Inquiry into the Critical Period for Human Speech* (Newbury House, New York).

Shiga, Y. (**2009**). "Pulse Density Representation of Spectrum for Statistical Speech Processing," in *Tenth Annual Conference of the International Speech Communication Association*.

Sidaras, S. K., Alexander, J. E., and Nygaard, L. C. (**2009**). "Perceptual learning of systematic variation in Spanish-accented speech," J Acoust. Soc. Am. **125**, 3306.

Silén, H., Helander, E., and Gabbouj, M. (**2011**). "Prediction of Voice Aperiodicity Based on Spectral Representations in HMM Speech Synthesis," in *Proceedings of INTERSPEECH*, pp. 105-108.

Sjöström, M., Eriksson, E. J., Zetterholm, E., and Sullivan, K. P. (**2006**). "A switch of dialect as disguise," Working Papers in Linguistics **52**, 113–116.

Stevens, K. N., Kasowski, S., and Fant, C. G. M. (**1953**). "An Electrical Analog of the Vocal Tract," J Acoust. Soc. Am. **25**, 734-742.

Stylianou, Y. (**2001**). "Applying the harmonic plus noise model in concatenative speech synthesis," IEEE Trans. on Speech and Audio Process. **9**, 21-29.

Stylianou, Y. (**2005**). "Modeling speech based on harmonic plus noise models," in *Nonlinear Speech Modeling and Applications* (Springer), pp. 244-260.

Stylianou, Y., Cappé, O., and Moulines, E. (**1998**). "Continuous probabilistic transform for voice conversion," IEEE Trans. Speech and Audio Process. **6**, 131-142.

Sullivan, K. P., and Schlichting, F. (**2007**). "Speaker discrimination in a foreign language: First language environment, second language learners," International Journal of Speech Language and the Law **7**, 95-112.

Sundermann, D., and Ney, H. (**2003**). "VTLN-based voice conversion," in *Proceedings of ISSPIT*, pp. 556-559.

Sundermann, D., Ney, H., and Hoge, H. (**2003**). "VTLN-based cross-language voice conversion," in *Proceedings of ASRU,* pp. 676-681.

Sundström, A. (**1998**). "Automatic prosody modification as a means for foreign language pronunciation training," in *Proceedings of ISCA Workshop on STiLL* pp. 49-52.

Takamichi, S., Toda, T., Black, A. W., and Nakamura, S. (**2014**). "Modulation spectrum-based post-filter for GMM-based Voice Conversion," in *Proceedings of APSIPA*, pp. 1-4.

Takashima, R., Takiguchi, T., and Ariki, Y. (**2012**). "Exemplar-based voice conversion in noisy environment," in *Proceedings of SLT,* pp. 313-317.

Tate, D. A. (**1979**). "Preliminary data on dialect in speech disguise," in *Proceedings of the IPS-77 congress*, pp. 847-850.

Tieleman, T. (**2008**). "Training restricted Boltzmann machines using approximations to the likelihood gradient," in *Proceedings of ICML*, pp. 1064-1071.

Toda, T., Black, A. W., and Tokuda, K. (**2004**). "Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis," in *Proceedings of ISCA SSW5*, pp. 31-36.

Toda, T., Black, A. W., and Tokuda, K. (**2007**). "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," IEEE Trans. Audio Speech Lang. Process. **15**, 2222-2235.

Toda, T., Black, A. W., and Tokuda, K. (**2008**). "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," Speech Commun. **50**, 215-227.

Toda, T., Muramatsu, T., and Banno, H. (**2012**). "Implementation of Computationally Efficient Real-Time Voice Conversion," in *Proceedings of INTERSPEECH*, pp. 94-97.

Torstensson, N., Eriksson, E. J., and Sullivan, K. P. (**2004**). "Mimicked accents. Do speakers have similar cognitive prototypes," in *Proceedings of the AICSST*, pp. 271-276.

Toth, A., and Black, A. (**2007**). "Using articulatory position data in voice transformation," Proceedings of ISCA SSW6, 182-187.

Toutios, A., and Maeda, S. (**2012**). "Articulatory VCV Synthesis from EMA Data," in *Proceedings of INTERSPEECH*, pp. 2566-2569.

Toutios, A., and Margaritis, K. (**2003**). "Acoustic-to-articulatory inversion of speech: a review," Proceedings of TAINN.

Toutios, A., and Narayanan, S. (**2013**). "Articulatory Synthesis of French Connected Speech from EMA Data," in *Interspeech*, pp. 2738-2742.

Traunmüller, H. (**1985**). "The role of the fundamental and the higher formants in the perception of speaker size, vocal effort, and vowel openness," in *Proceedings of Franco-Swedish Seminar on Speech* (Grenoble, France), pp. 209-219.

Traunmüller, H. (**1994**). "Conventional, biological and environmental factors in speech communication: a modulation theory," Phonetica **51**, 170-183.

Traunmüller, H. (**2005**). "Speech considered as modulated voice."

Turk, O., and Arslan, L. M. (**2006**). "Robust processing techniques for voice conversion," Computer Speech & Language **20**, 441-467.

Uria, B., Murray, I., Renals, S., and Richmond, K. (**2012**). "Deep architectures for articulatory inversion," in *Proceedings of INTERSPEECH*, pp. 867-870.

Wade, T., Jongman, A., and Sereno, J. (**2007**). "Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds," Phonetica **64**, 122-144.

Westbury, J. R. (**1994**). *X-Ray microbeam speech production database user's handbook version 1.0* (Waisman Center on Mental Retardation & Human Development, University of Wisconsin, Madison, WI).

Wrench, A. (**1999**). "MOCHA-TIMIT," (Queen Margaret University College).

Xingyu, N., Xiang, X., and Jingming, K. (**2014**). "Low latency parameter generation for real-time speech synthesis system," in *Proceedings of ICME*, pp. 1-6.

Yan, Q., Vaseghi, S., Rentzos, D., and Ho, C.-H. (**2004**). "Analysis by synthesis of acoustic correlates of British, Australian and American accents," in *Proceedings of ICASSP*, pp. I-637-640.

Yan, Q., Vaseghi, S., Rentzos, D., and Ho, C.-H. (**2007**). "Analysis and synthesis of formant spaces of British, Australian, and American accents," IEEE Trans. Audio, Speech, and Lang. Process. **15**, 676-689.

You, H., Alwan, A., Kazemzadeh, A., and Narayanan, S. (**2005**). "Pronunciation variations of Spanish-accented English spoken by young children," in *Proceddings of INTERSPEECH*, pp. 749-752.

Young, S. (**1996**). "A review of large-vocabulary continuous-speech," Signal Processing Magazine **13**, 45.

Youssef, A. B., Hueber, T., Badin, P., and Bailly, G. (**2011**). "Toward a multi-speaker visual articulatory feedback system," in *Proceedings of INTERSPEECH*, pp. 589-592.

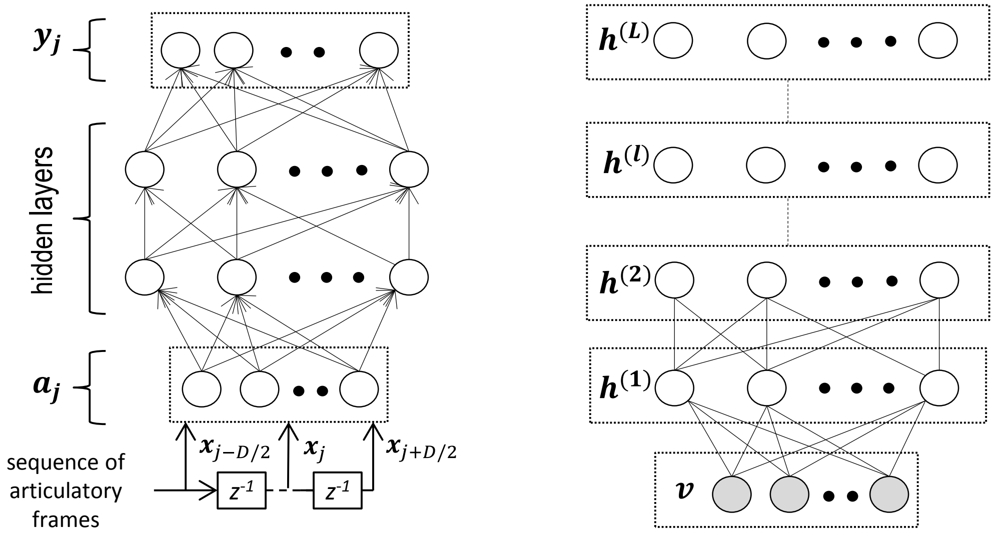
Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., and Tokuda, K. (**2007**). "The HMM-based speech synthesis system (HTS) version 2.0," in *Proceedings of ISCA SSW*, pp. 294-299.

Zen, H., Senior, A., and Schuster, M. (**2013**). "Statistical parametric speech synthesis using deep neural networks," in *Proceedings of ICASSP*, pp. 7962-7966.

APPENDIX A: FORWARD MAPPING WITH DEEP NETWORKS

As illustrated in Figure 36a, the DNN consists of an input layer, an output layer, and multiple layers of hidden units between them. In this particular topology, units in a layer are fully connected to units in the immediate layer above it, but there is no connection among units within a layer. The network contains a tapped-delay line at the input that allows the model to consider not only the current articulatory configuration $\boldsymbol{x}_j$ but also that of nearby frames, resulting in the input vector $\boldsymbol{a}_j = \{\boldsymbol{x}_{j-D/2} \dots \boldsymbol{x}_j \dots \boldsymbol{x}_{j+D/2}\}$, where D is the number of delay units in the tapped-delay line. When $D = 0$, the input vector $\boldsymbol{a}_j$ becomes the articulatory feature vector $\boldsymbol{x}_j$, and the DNN performs a *frame-by-frame* mapping. Increasing the value of D allows the DNN to include additional temporal context to aid in predicting the acoustic observation $\boldsymbol{y}_j$.

We train the DNN using the conventional two-stage hybrid recipe (Hinton, 2012). During the first stage, model parameters (for all but the last layer) are learned in an unsupervised fashion; this pre-training stage makes it more likely to find a good local optimum than using randomly initialized parameters (Erhan *et al.*, 2010). During the second stage, the pre-trained model (including the last layer) is fine-tuned in a supervised fashion via back-propagation (Rumelhart *et al.*, 1986).



(a) (b)

Figure 36: (a) Forward mapping via a deep neural network (DNN) with a tapped-delay line input. (b) The Gaussian-Bernoulli deep Boltzmann machine as an undirected graphical model with real valued visible units $\boldsymbol{v}$ and binary hidden units $\boldsymbol{h}$.

Pre-training the network as a generative model

During pre-training the network is operated as a Gaussian-Bernoulli deep Boltzmann machine (GDBM), an energy-based generative model that allows each unit to receive both top-down and bottom-up signals (Cho *et al.*, 2013). Unlike a generic deep Boltzmann machine (Salakhutdinov and Hinton, 2009), which has binary units in all of its layers, the GDBM has Gaussian units in the visible layer, making it better suited to handle real-valued inputs.

Consider the simplified GDBM with multiple hidden layers shown in Figure 36b, where $\boldsymbol{v}$ represents a visible layer of real valued input variables, and $\{\boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)} \ldots \boldsymbol{h}^{(l)} \ldots \boldsymbol{h}^{(L)}\}$ represent the L hidden layers of binary variables. Following (Cho *et al.*, 2013), the energy of the GDBM at state $\{\boldsymbol{v}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)} .. \boldsymbol{h}^{(L)}\}$ is defined by:

$$E\left(\boldsymbol{v}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)} \dots \boldsymbol{h}^{(l)} \dots \boldsymbol{h}^{(L)} \middle| \boldsymbol{\theta}\right) = \sum_{i=1}^{N^{(v)}} \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_{i=1}^{N^{(v)}} \sum_{j=1}^{N^{(1)}} \frac{v_i}{2\sigma_i^2} h_j^{(1)} w_{ij} - \sum_{l=1}^{L} \sum_{j=1}^{N^{(l)}} b_j^{(l)} h_j^{(l)} - \sum_{l=1}^{L-1} \sum_{j=1}^{N^{(l)}} \sum_{k=1}^{N^{(l+1)}} h_j^{(l)} h_k^{(l+1)} w_{jk}^{(l)}, \tag{31}$$

where $\boldsymbol{\theta} = \{\boldsymbol{b}, \boldsymbol{\sigma}, \boldsymbol{b}^{(l)}, \boldsymbol{w}, \boldsymbol{w}^{(l)}\}$ is the set of model parameters consisting of biases $\boldsymbol{b}$ and standard deviations $\boldsymbol{\sigma}$ for visible units; biases $\boldsymbol{b}^{(l)}$ for hidden layer $\boldsymbol{l}$; weights $\boldsymbol{w}$ for the connections between the visible and the first hidden layer, and weights $\boldsymbol{w}^{(l)}$ for the connections between the units in the l-th and l+1-th hidden layer. $N^{(v)}$and $N^{(l)}$ are the number of units in the visible layer and the l-th hidden layer, respectively. Given this energy function, the corresponding probability for state $\{\boldsymbol{v}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}, \dots \boldsymbol{h}^{(l)} \dots \boldsymbol{h}^{(L)}\}$ can be computed as:

$$p\left(\boldsymbol{v}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)},, \dots \boldsymbol{h}^{(l)} \dots \boldsymbol{h}^{(L)} \middle| \boldsymbol{\theta}\right) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(-E(\boldsymbol{v}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}, \dots \boldsymbol{h}^{(l)} \dots \boldsymbol{h}^{(L)} | \boldsymbol{\theta}\,)\right) \tag{32}$$

where $Z(\boldsymbol{\theta})$ is the normalizing factor over all possible values of $\{\boldsymbol{v}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}, \dots \boldsymbol{h}^{(l)} \dots \boldsymbol{h}^{(L)}\}$. This relation between the probability and the energy of a state is designed such that the model rarely reaches high energy states. The conditional probability of the input variable, v_{i} (a Gaussian unit), given the hidden units $\boldsymbol{h}^{(1)}$ is defined by the normal distribution:

$$p\left(v_{\mathrm{i}}\middle|\boldsymbol{h}^{(1)},\boldsymbol{\theta}\right)=\mathcal{N}\left(v_i\middle|\sum\nolimits_{j=1}^{N^{(1)}}h_j^{(1)}w_{ij}+b_i\ ,\sigma_i^2\right),i=1\colon N^{(v)} \tag{33}$$

whereas the conditional probabilities of hidden units are defined by

$$p\left(h_{\mathrm{j}}^{(1)}\middle|\boldsymbol{v},\boldsymbol{h}^{(2)},\boldsymbol{\theta}\right)=sig\left(\sum\nolimits_{i=1}^{N^{(v)}}\frac{v_i}{2\sigma_i^2}w_{ij}+\sum\nolimits_{k=1}^{N^{(2)}}h_k^{(2)}w_{jk}^{(1)}+b_j^{(1)}\right), \tag{34}$$
$$j=1\colon N^{(1)}$$

and

$$p\left(h_{\mathrm{j}}^{(l)}\middle|\boldsymbol{h}^{(l-1)},\boldsymbol{h}^{(l+1)},\boldsymbol{\theta}\right)$$
$$=sig\left(\sum\nolimits_{i=1}^{N^{(l-1)}}h_i^{(l-1)}w_{ij}^{(l-1)}+\sum\nolimits_{k=1}^{N^{(\mathrm{l}+1)}}h_k^{(l+1)}w_{jk}^{(l)}+b_j^{(l)}\right) \tag{35}$$
$$l>1\ and\ j=1\colon N^{(l)}$$

where $sig(.)$ represents the logistic sigmoid function $sig(z)=1/(1+e^{-z})$.

Thus, a GDBM can be thought of as a generative graphical model whose conditional probabilities, defined by equations (33)-(35), depend on a set of model parameters $\boldsymbol{\theta}$. These model parameters are then trained such that the graphical model represents the distribution of input vectors in the training set. Namely, being an energy-based model, the GDBM is trained to reduce the energy of configurations in the training data and increase the energy of any other configurations that could be generated by the model. The energy of a configuration is related to its probability, as given in equation (32). Thus, the training process is equivalent to performing stochastic gradient ascent on the log-likelihood of the training data, which can be shown (Hinton and Sejnowski, 1983) to lead to the parameter update equations:

$$\Delta\boldsymbol{w} = \alpha\left(\mathbb{E}_{pData}\left\{\boldsymbol{v}.\boldsymbol{h}^{(\mathbf{1})^{\top}}\right\} - \mathbb{E}_{pModel}\left\{\boldsymbol{v}.\boldsymbol{h}^{(\mathbf{1})^{\top}}\right\}\right) \tag{36}$$

$$\Delta\boldsymbol{w}^{(\boldsymbol{l})} = \alpha\left(\mathbb{E}_{pData}\left\{\boldsymbol{h}^{(\boldsymbol{l})}.\boldsymbol{h}^{(\boldsymbol{l}+\mathbf{1})^{\top}}\right\} - \mathbb{E}_{pModel}\left\{\boldsymbol{h}^{(\boldsymbol{l})}.\boldsymbol{h}^{(\boldsymbol{l}+\mathbf{1})^{\top}}\right\}\right) \text{ for } l \geq 1 \tag{37}$$

$$\Delta\boldsymbol{b} = \alpha\left(\mathbb{E}_{pData}\{\boldsymbol{v}\} - \mathbb{E}_{pModel}\{\boldsymbol{v}\}\right) \tag{38}$$

$$\Delta\boldsymbol{b}^{(\boldsymbol{l})} = \alpha\left(\mathbb{E}_{pData}\left\{\boldsymbol{h}^{(\boldsymbol{l})}\right\} - \mathbb{E}_{pModel}\left\{\boldsymbol{h}^{(\boldsymbol{l})}\right\}\right) \text{ for } l \geq 1 \tag{39}$$

where $\mathbb{E}_{pData}$ is the data-dependent expectation, calculated over the conditional distribution $p(\boldsymbol{h}|\boldsymbol{v}\{tr\},\boldsymbol{\theta})$, and $\boldsymbol{v}\{tr\}$ is a sample in the training set. In contrast, $\mathbb{E}_{pModel}$ is the model's expectation, calculated over the distribution $p(\boldsymbol{h},\boldsymbol{v}|\boldsymbol{\theta})$. Exact calculation of these expectations is intractable because the time required grows exponentially with the number of hidden units –see equations (33)-(35). Fortunately, practical approximations of these expectations are possible. In particular, we use the mean-field approximation to calculate data-dependent expectation (Salakhutdinov and Hinton, 2009) and a variation of Markov-chain Monte-Carlo sampling to calculate model's expectation (Cho *et al.*, 2013).

In the mean-field approximation of the expectation over the data distribution, visible units are first clamped to a training sample. Then, the hidden units are described as having the probability $\boldsymbol{\mu}$ of being active, which is iteratively updated until convergence with the fixed-point iteration:

$$\mu_j^{(l)} \leftarrow sigm\left(\sum_{i=1}^{N^{(l-1)}} \mu_i^{(l-1)} w_{ij}^{(l-1)} + \sum_{k=1}^{N^{(l+1)}} \mu_k^{(l+1)} w_{jk}^{(l)} + b_j^{(l)}\right);\ l \geq 1 \tag{40}$$

with boundary conditions given by $\mu_i^{(0)} = v_i/\sigma_i^2$ and $N^{(0)} = N^{(v)}$. For each sample $\boldsymbol{v}\{tr\}$ in the training set, we calculate the corresponding set of $\boldsymbol{\mu}$. The data-dependent expectations in equations (36)-(39) are then approximated by replacing $\boldsymbol{h}$ by the corresponding $\boldsymbol{\mu}$ and then averaging over all the training samples.

The model's expectancy is calculated using a Markov-chain Monte-Carlo method (MCMC). Given the conditional probability distributions, we sample both visible and hidden variables using MCMC and use these samples to calculate the model expectations in equations (36)-(39). Specifically, we use the parallel tempering approach of Cho et al. (Cho *et al.*, 2013) to maintain multiple persistent Gibbs sampling chains. The persistent chains (Tieleman, 2008) help reduce the computational cost by updating only a few samples from each chain at each model parameter update instead of starting a new chain. Usually, modifying only a few samples in the chains is sufficient to represent the updated model because the updates in parameters are too small to make a significant change in the probability distribution. However, in cases where the model distribution changes significantly, persistent chains may not be able to evolve to represent the updated probability distribution. The parallel tempering approach helps alleviate this problem by maintaining multiple chains at different temperatures. In high temperature chains samples are more likely to explore the state space, whereas in low temperature chains samples follow the target model distribution.

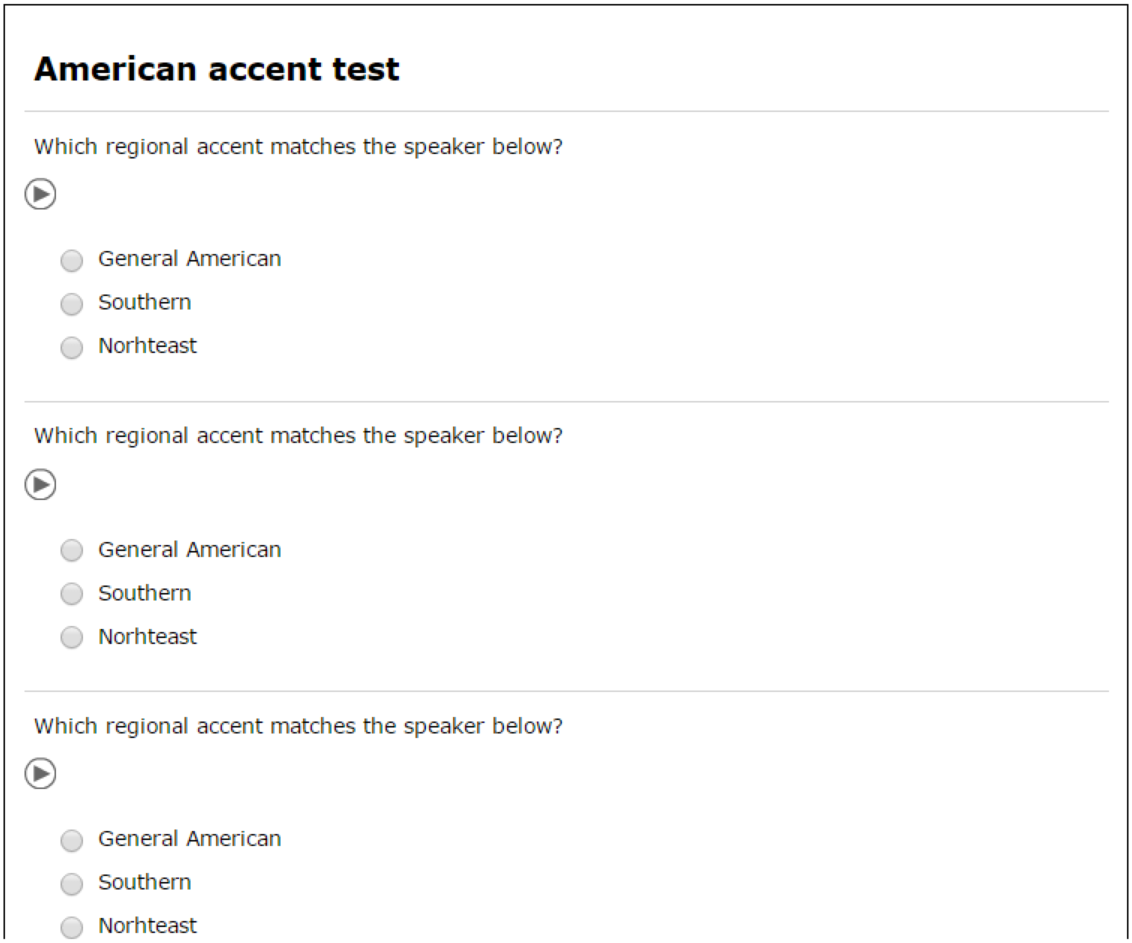
Training a GDBM as described above is a slow learning process, particularly for hidden layers remote from the visible units. We speed up the process by following the greedy layer-wise method commonly used for deep Boltzmann machines (Salakhutdinov and Hinton, 2009).

Building a DNN from a trained GDBM

Once the underlying GDBM is trained, we build a DNN as follows. First, a layer of output units is added to the topmost hidden layer of the GDBM, one output unit for each corresponding acoustic feature. Connection weights between the units at the topmost hidden layer and the newly added output layer are initialized randomly. The resulting multilayer neural network is then discriminatively fine-tuned using standard back-propagation (Rumelhart *et al.*, 1986).

APPENDIX B: MECHANICAL TURK TESTS SAMPLES

In this appendix, we have listed samples of the web-based forms used in subjective evaluation of foreign accent conversions. We collected the listeners' responses using these forms via an online crowdsourcing platform hosted by Amazon services, Mechanical Turk. These samples are (i) the qualification task asking participants to classify the American Accents, (ii) a typical intelligibility evaluation task, (iii) a forced pairwise comparison of perceived non-native accentedness, (iv) a typical voice-similarity evaluation task, and (v) a task for the subjective evaluation of acoustic quality.



American accent test

Which regional accent matches the speaker below?

○ General American
○ Southern
○ Norhteast

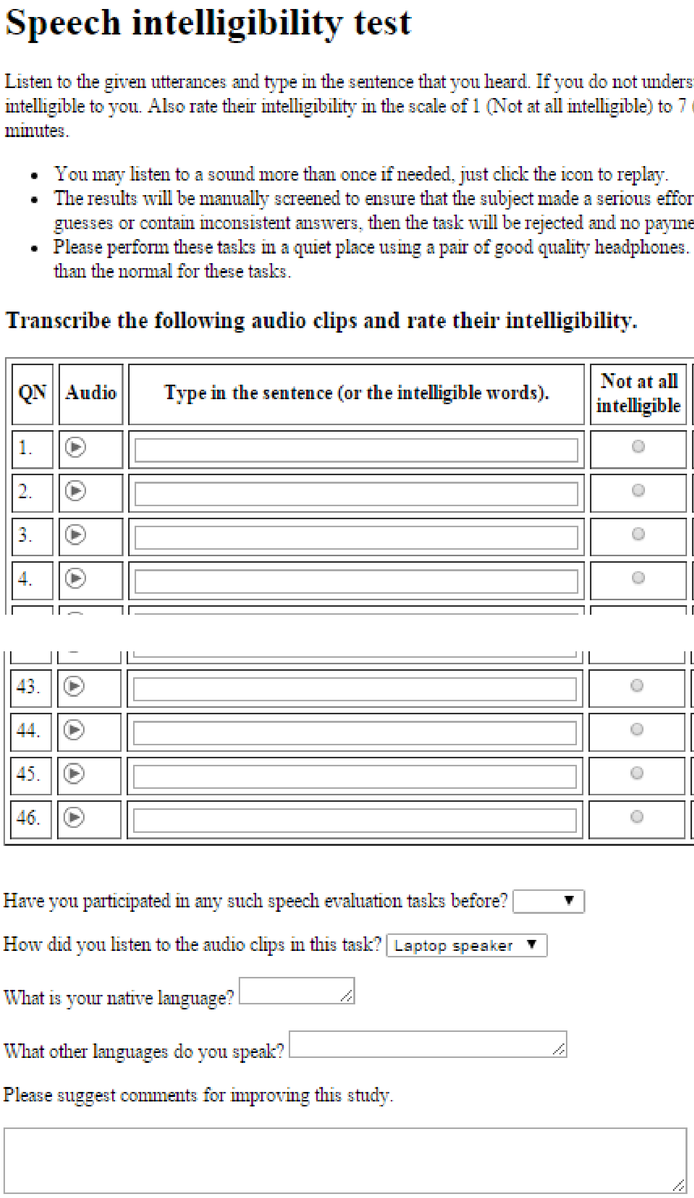
Which regional accent matches the speaker below?

○ General American
○ Southern
○ Norhteast

Which regional accent matches the speaker below?

○ General American
○ Southern
○ Norhteast

Figure 37: The pre-qualification test to identify the American accents. This qualification test was used to select the native speakers of American English for the perceptual listening tests.



Speech intelligibility test

Listen to the given utterances and type in the sentence that you heard. If you do not understand the complete sentence, type in the words that are intelligible to you. Also rate their intelligibility in the scale of 1 (Not at all intelligible) to 7 (Extremely intelligible). This test should take about 30 minutes.

- You may listen to a sound more than once if needed, just click the icon to replay.
- The results will be manually screened to ensure that the subject made a serious effort to complete the task. If the results are determined to be guesses or contain inconsistent answers, then the task will be rejected and no payment will be given.
- Please perform these tasks in a quiet place using a pair of good quality headphones. It is recommended that you set the volume a bit higher than the normal for these tasks.

Transcribe the following audio clips and rate their intelligibility.

QN	Audio	Type in the sentence (or the intelligible words).	Not at all intelligible	|	Somewhat intelligible	|	Quite a bit intelligible	|	Extremely intelligible
1.	▶		○	○	○	○	○	○	○
2.	▶		○	○	○	○	○	○	○
3.	▶		○	○	○	○	○	○	○
4.	▶		○	○	○	○	○	○	○
43.	▶		○	○	○	○	○	○	○
44.	▶		○	○	○	○	○	○	○
45.	▶		○	○	○	○	○	○	○
46.	▶		○	○	○	○	○	○	○

Have you participated in any such speech evaluation tasks before? ▼

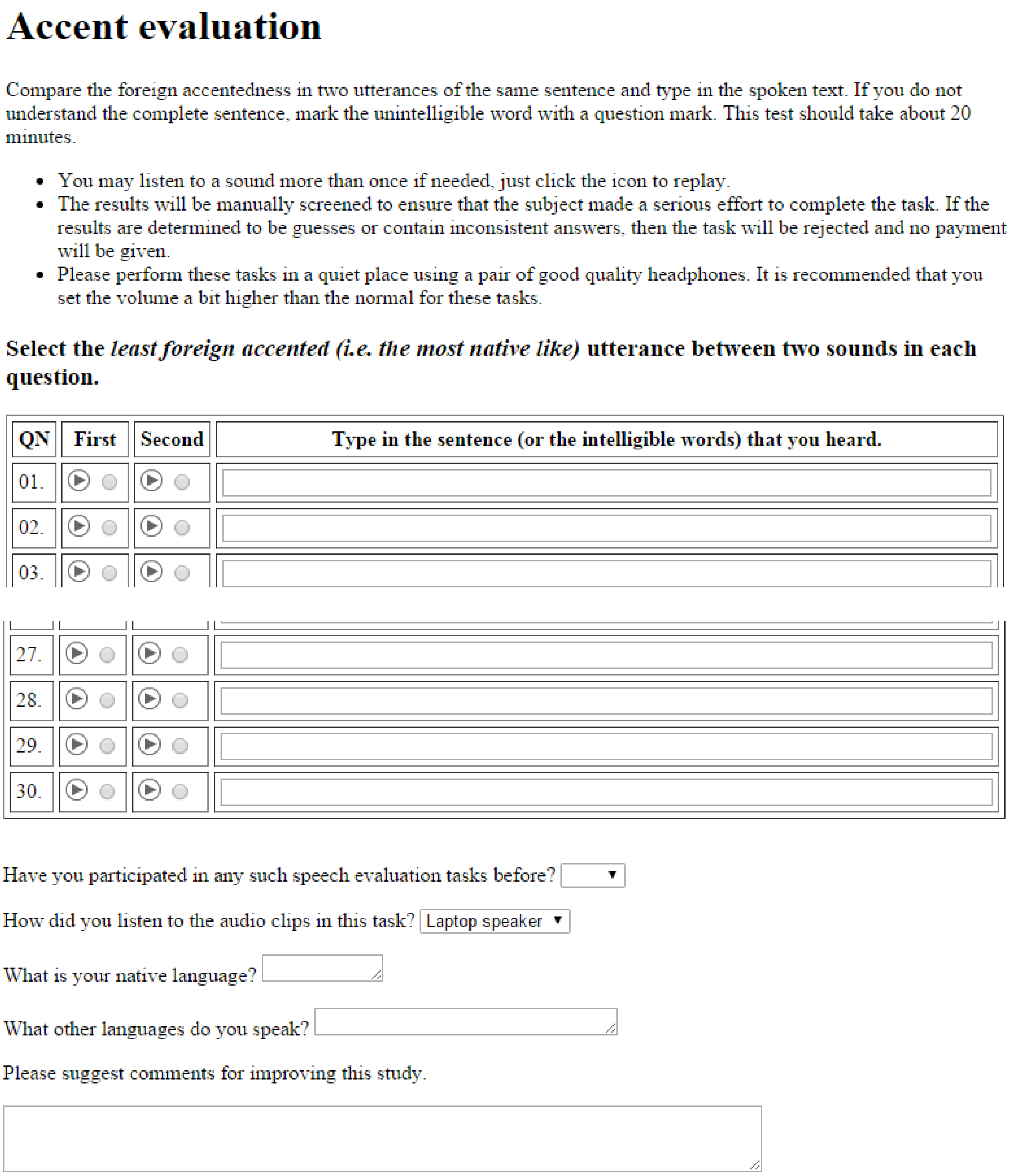
How did you listen to the audio clips in this task? Laptop speaker ▼

What is your native language?

What other languages do you speak?

Please suggest comments for improving this study.

Figure 38: The intelligibility assessment test. The participants were asked to transcribe the audio clips and rate their intelligibility.



Accent evaluation

Compare the foreign accentedness in two utterances of the same sentence and type in the spoken text. If you do not understand the complete sentence, mark the unintelligible word with a question mark. This test should take about 20 minutes.

- You may listen to a sound more than once if needed, just click the icon to replay.
- The results will be manually screened to ensure that the subject made a serious effort to complete the task. If the results are determined to be guesses or contain inconsistent answers, then the task will be rejected and no payment will be given.
- Please perform these tasks in a quiet place using a pair of good quality headphones. It is recommended that you set the volume a bit higher than the normal for these tasks.

Select the *least foreign accented (i.e. the most native like)* utterance between two sounds in each question.

QN	First	Second	Type in the sentence (or the intelligible words) that you heard.
01.	▶ ○	▶ ○	
02.	▶ ○	▶ ○	
03.	▶ ○	▶ ○	
27.	▶ ○	▶ ○	
28.	▶ ○	▶ ○	
29.	▶ ○	▶ ○	
30.	▶ ○	▶ ○	

Have you participated in any such speech evaluation tasks before? ▼

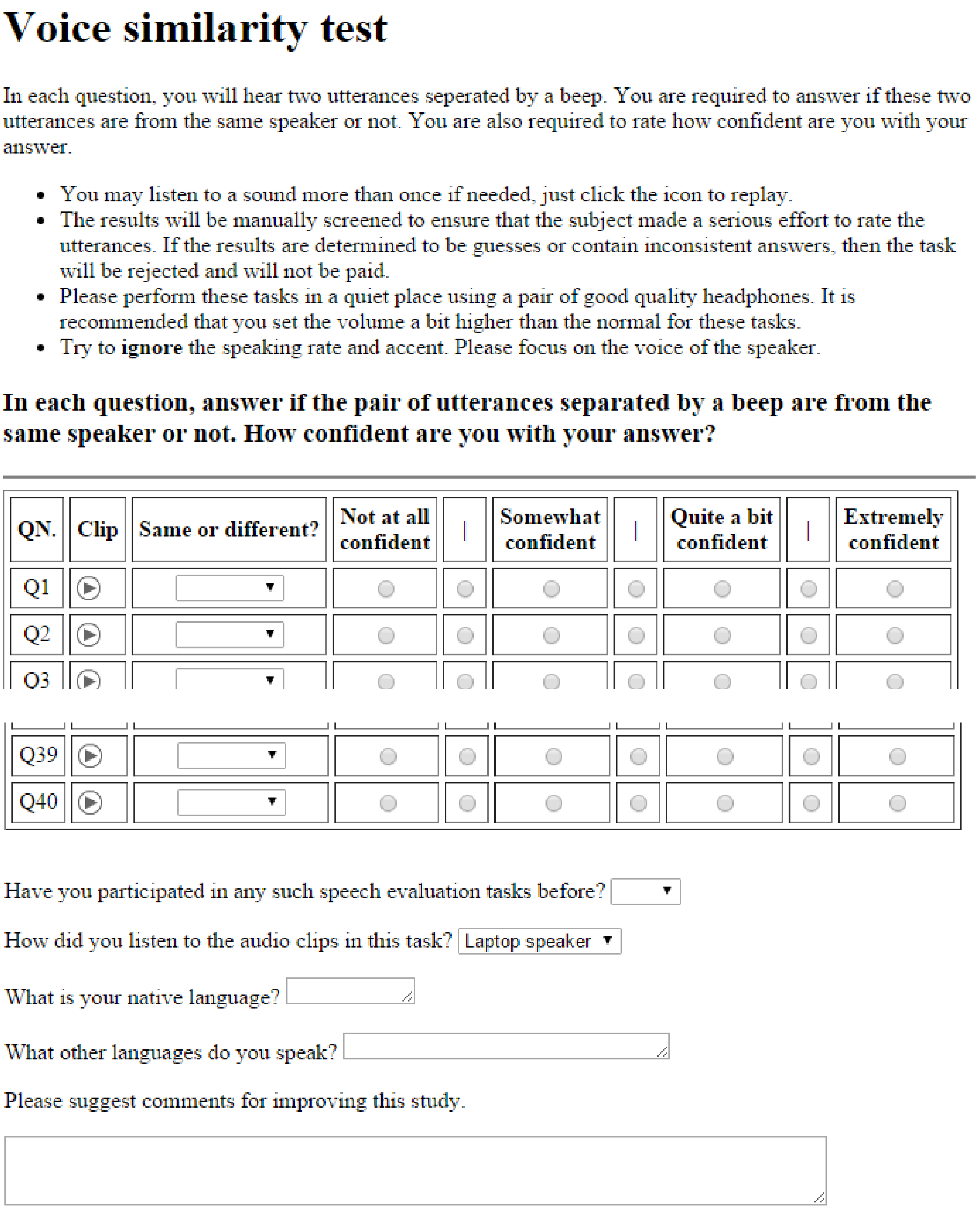
How did you listen to the audio clips in this task? Laptop speaker ▼

What is your native language?

What other languages do you speak?

Please suggest comments for improving this study.

Figure 39: Forced pairwise comparison of non-native accentedness. The participants listened to the two clips of the same sentence and selected the one that sounded the most native-like. Participants were also asked to transcribe the sentence to ensure that they listen to the sentence.



Voice similarity test

In each question, you will hear two utterances seperated by a beep. You are required to answer if these two utterances are from the same speaker or not. You are also required to rate how confident are you with your answer.

- You may listen to a sound more than once if needed, just click the icon to replay.
- The results will be manually screened to ensure that the subject made a serious effort to rate the utterances. If the results are determined to be guesses or contain inconsistent answers, then the task will be rejected and will not be paid.
- Please perform these tasks in a quiet place using a pair of good quality headphones. It is recommended that you set the volume a bit higher than the normal for these tasks.
- Try to **ignore** the speaking rate and accent. Please focus on the voice of the speaker.

In each question, answer if the pair of utterances separated by a beep are from the same speaker or not. How confident are you with your answer?

QN.	Clip	Same or different?	Not at all confident	|	Somewhat confident	|	Quite a bit confident	|	Extremely confident
Q1	▶	▼	○	○	○	○	○	○	○
Q2	▶	▼	○	○	○	○	○	○	○
Q3	▶	▼	○	○	○	○	○	○	○
Q39	▶	▼	○	○	○	○	○	○	○
Q40	▶	▼	○	○	○	○	○	○	○

Have you participated in any such speech evaluation tasks before? ▼

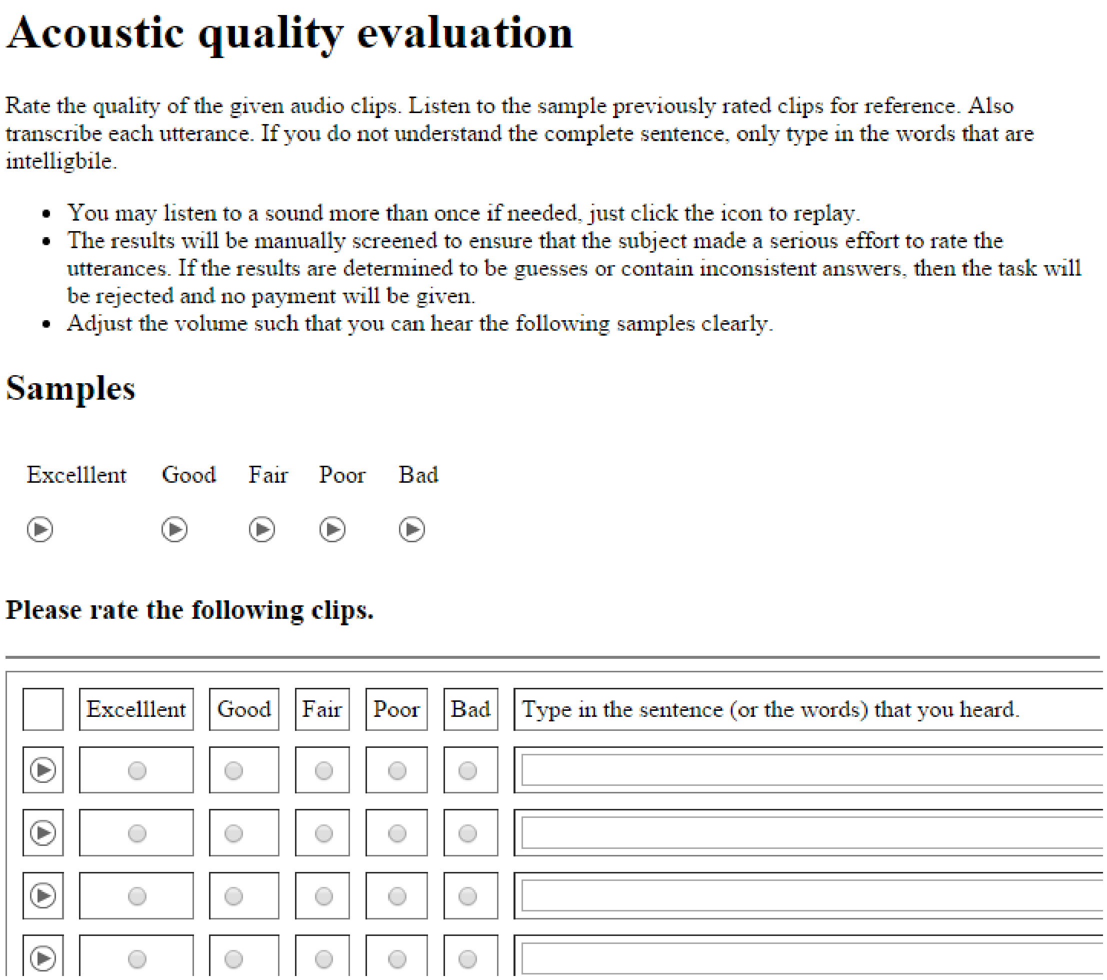
How did you listen to the audio clips in this task? Laptop speaker ▼

What is your native language?

What other languages do you speak?

Please suggest comments for improving this study.

Figure 40: The voice-similarity test asks the participants to listen to a pair of linguistically different utterances separated by a beep and answer (i) if the utterances were from the same speaker or not, and (ii) how confident they are on their decision.



Acoustic quality evaluation

Rate the quality of the given audio clips. Listen to the sample previously rated clips for reference. Also transcribe each utterance. If you do not understand the complete sentence, only type in the words that are intelligbile.

- You may listen to a sound more than once if needed, just click the icon to replay.
- The results will be manually screened to ensure that the subject made a serious effort to rate the utterances. If the results are determined to be guesses or contain inconsistent answers, then the task will be rejected and no payment will be given.
- Adjust the volume such that you can hear the following samples clearly.

Samples

Excelllent	Good	Fair	Poor	Bad
▶	▶	▶	▶	▶

Please rate the following clips.

	Excelllent	Good	Fair	Poor	Bad	Type in the sentence (or the words) that you heard.
▶	○	○	○	○	○	
▶	○	○	○	○	○	
▶	○	○	○	○	○	
▶	○	○	○	○	○	

Figure 41: Subjective evaluation of the acoustic quality. The participants were asked to rate the utterances in the MOS scale (Bad:1, Poor:2, Fair:3 Good:4, Excellent:5). Rated samples were provided for the reference.

APPENDIX C: LIST OF PUBLICATIONS

Following is the list of publications related to this dissertation work.

Journal articles:

1. Aryal, S., and Gutierrez-Osuna, R. (2015). "Reduction of non-native accents through statistical parametric articulatory synthesis," J. Acoust. Soc. Am. 137, 433-446.
2. Aryal, S., and Gutierrez-Osuna, R. (2015). "*Data driven articulatory synthesis with deep neural networks,*" Computer Speech & Language.

Conference proceedings:

1. Aryal, S., and Gutierrez-Osuna, R. (2015). "*Articulatory-based conversion of foreign accents with deep neural networks*," in Proceedings of INTERSPEECH, pp. 3385-3389.
2. Christopher Liberatore, Sandesh Aryal, Zelun Wang, Seth Polsley and Ricardo Gutierrez-Osuna (2015). "*SABR: Sparse, Anchor-Based Representation of the Speech Signal,*" in Proceedings of INTERSPEECH, pp. 608-612
3. Aryal, S., and Gutierrez-Osuna, R. (2014). "*Can voice conversion be used to reduce non-native accents?*," in Proceedings of ICASSP, pp. 7929-7933.
4. Felps, D., Aryal, S., and Gutierrez-Osuna, R. (2014). "*Normalization of articulatory data through Procrustes transformations and analysis-by-synthesis*," in Proceedings of ICASSP, pp. 3051-3055.

5. Aryal, S., and Gutierrez-Osuna, R. (2014). "*Accent conversion through cross-speaker articulatory synthesis*," in Proceedings of ICASSP, pp. 7744-7748.
6. Tian Lan, Sandesh Aryal, Beena Ahmed, Kirrie Ballard, and Ricardo Gutierrez-Osuna (2014). "*Flappy Voice: An Interactive Game for Childhood Apraxia of Speech Therapy*," in Proceedings of CHI-PLAY, pp. 429-430.
7. Aryal, S., and Gutierrez-Osuna, R. (2013). "*Articulatory inversion and synthesis: towards articulatory-based modification of speech*," in Proceedings of ICASSP, pp. 7952-7956.
8. Aryal, S., Felps, D., and Gutierrez-Osuna, R. (2013). "*Foreign accent conversion through voice morphing*," in Proceedings of INTERSPEECH, pp. 3077-3081.

APPENDIX D: PSI STATFAC TOOLBOX

The PSI StatFAC toolbox is a Matlab library developed at the Texas A&M University PSI lab. The library provides the Matlab functions and scripts required to perform foreign accent conversion (FAC) based on statistical parametric approaches developed in this dissertation work. The library was developed in Matlab (2012b) and it requires an in-house toolbox ConFAC (Felps *et al.*, 2012), and third party toolboxes, namely, DEEPMAT (Cho, 2013), STRAIGTH (Kawahara, 1997), and NETLAB (Nabney, 2002).

INSTALLATION AND OVERVIEW

Installation involves copying the library folder and adding all its directories and subdirectories to the Matlab path. The toolbox contains the four main subdirectories: (i) *exampleScripts*, which contains the sample scripts to load and preprocess training data, (ii) *acsout_based*, which contains the functions required for acoustic-based accent conversion, (iii) *art_based*, which contains the functions required for articulatory-based foreign accent conversion, and (iv) *thirdpartytoolboxes*, which contains the required third-party toolboxes. All the required third-party toolboxes are already included in the StatFAC distribution. To install ConFAC, please refer to the ConFAC user manual for instructions.

For acoustic-based accent conversion, the toolbox provides all the functionalities needed to perform FAC given a speech corpus from a native (L1) and a non-native (L2) speaker. The corpus should contain the audio ('.wav') and the transcription files ('.lab') for all the utterances organized in two folders, one each for L1 and L2. A sample dataset is provided in the subdirectory *testData*.

In case of the articulatory-based accent conversion, the toolbox relies on ConFAC to access the corpus, to extract features, and to generate waveform. ConFAC also comes with the parallel speech corpus from two speakers, a native speaker of American English (MAB) and a native speaker of Spanish (RGO). The corpus is used in the examples provided with this manual to illustrate the StatFAC functionalities available for articulatory-based accent conversion. The corpus contains the audio recordings, articulatory recordings collected via electromagnetic articulography (EMA), and the phonetic transcriptions[20] –see (Felps, 2011) for more details on the corpus. In order to use a corpus from a new speaker, please refer to ConFAC user manual for instructions on adding a new speaker.

EXAMPLES

In this document, we describe the accent conversion functionalities (e.g., training the models, estimating acoustic features) and the auxiliary functionalities (e.g., loading corpus, extracting features, and generating waveforms) available in StatFAC using examples. These examples illustrate the processes involved in two different foreign accent conversion methods supported by StatFAC. First, we describe the acoustic-based conversion using the audio recordings from a given pair of native and non-native

[20] Phonetic transcriptions provide the nasality feature in articulatory-based methods.

speaker. Secondly, we describe the articulatory-based methods using the corpus provided with the ConFAC distribution as an example.

Acoustic-based foreign accent conversion

We now explain the steps required to perform the acoustic-based accent conversion using the cross-speaker statistical mapping. The overall process consists of six steps. In the first step, we load the speech corpus from the L1 and L2 speakers, and prepare the training dataset. In the second step, we extract global variances (GVs) from the training utterances for each speaker. In the third step, we extract the mean and variance of $log(f_0)$ for both the speakers so that it can be used in modifying L1 pitch trajectories to match the range of the L2 speaker. In the fourth step, we pair the acoustic feature vectors from the L1 speaker with that of the L2 speaker; these pairings are used to train cross-speaker statistical mapping. For accent conversion, we pair the frames based on their acoustic similarity. StatFAC also provides a function to force-align the parallel utterances. The force-aligned frames are used to train the mappings for conventional voice conversion. In the fourth step, we train a GMM-based cross-speaker mapping model. Finally, we use the model to generate speech from a test L1 utterance. We now describe these five steps with examples.

Step 1: Prepare the training dataset

The data preparation step takes the audio recordings from the native and non-native speaker, extracts acoustic features (MFCCs), and creates training datasets as required for the subsequent steps. A sample script for data preparation (*exampleScripts/dataPreparationAcoustMethod.m)* is provided with the toolbox. StatFAC requires the audio ('.wav') and the transcription files ('.lab') for all the utterances from the native (L1) and non-native (L2) speaker to be organized in two folders, one for each speaker. The audio and the transcription file for the same utterance should share the same name with '.wav' and '.lab' extensions, respectively. Furthermore, the label file should be in HTK format with the phoneme labels based on CMU pronunciation dictionary. The script extracts STRAIGHT parameters for each audio file. From the STRAIGHT spectrogram, the script, then, extracts the acoustic feature vector for each frame sampled at 200Hz. The extracted parameters for an utterance are saves as Matlab data file with the same filename but with the extension '.mat' in the same folder.

For the training purpose, the script requires a list of training utterances from the two speakers. The list is a text file containing the filenames of the parallel utterances (without '.wav' extension) in two columns. The first column contains the filenames of the training utterances from L1 and the second column contains the filenames of the same sentence for L2. A typical list of training utterances is available in the subdirectory *testData*. Given such a list, for each speaker, the script assembles the non-silent frames (in sequential order) from the training utterances and stores in variables `tr_MFCC`, `tr_logf0` `tr_uniqPhLblsID`, and `tr_utt_id`; each row of these variables respectively corresponds to the $MFCC_{1-24}$ (excluding the frame energy $MFCC_0$) and their derivatives, log of fundamental frequency ($logf_0$), phone label of the frame, and the index to the utterance to which the frame belongs. The index to the utterance is required to identify the pair of parallel utterances from the two speakers while performing forced-alignment. These variables are saved for subsequent use. In the given example script, the variables corresponding to the training utterances for L2 are saved in `trDataset_L2.mat`, whereas, the variables for L1 are saved in `trDataset_L1.mat`.

Step 2: Extract Global Variance (GV)

We extract the GVs for the L2 speaker using only the training utterances. The example below shows the commands to load the training data and calculate the GVs for each utterance in the training set. We save the extracted GVs for future use.

```
% Load the training data for the L2 speakers
L2TrData = load('trDataset_L2');
[trUttGVs] = calculateGVs(L2TrData);
save trUttGVsL2.mat trUttGVs;
```

Step 3: Calculate parameters for pitch-modification

We now calculate the pitch-modification parameters (`pitchXForm`) that is required to convert pitch trajectory from the L1 utterances to match the range of the L2 speaker.

```
L2TrData = load('trDataset_L2');
L1TrData = load('trDataset_L1');

% pitch transformation from L1 to L2 pitch range
isInputPitch= 1; % input is pitch not the articulatory feature
vectors
[pitchXForm] = calculatePitchTransform(L1TrData.tr_logf0,
L2TrData.tr_logf0, isInputPitch);
save pitchXForm_L1toL2 pitchXForm;
```

Step 4: Pair frames from L1 and L2

In this step, the frames from L1 and L2 training utterances are paired with each other to create a pairs of acoustic feature vectors. These pairings are used model the joint probability distribution of the acoustic feature vectors from both the speakers using Gaussian mixtures. StatFAC offers two types of pairing techniques: (i) force aligned pairing of the parallel utterances using dynamic time warping (DTW), and (ii) pairing based on the acoustic similarity between the L1 and L2 frames, following vocal tract length normalization (VTLN). While the former pairing technique leads to the conventional voice conversion, the latter leads to accent conversion. The example script below generates pairs of frames from L2 and L1 training utterances. Sample commands for both the pairing techniques are given.

```
% Load the preprocessed training data for the two speakers,
L1TrData = load('trDataset_L1');
L2TrData = load('trDataset_L2');
% forced time-aligned pairing of parallel utterances using DTW
[L1MFCC_FA, L2MFCC_FA] = framePairing_dtw(L1TrData, L2TrData);
% acsoutic similarity based frame pairing between two speakers
[L1MFCC_AcSim,L2MFCC_AcSim]=framePairing_acSim(L1TrData,L2TrData);
```

Step 5: Train the cross-speaker statistical mapping

We train the cross-speaker statistical mappings from the L1 speaker to the L2 using a Gaussian mixture model (GMM). Given the set of frame-pairs from both the speakers, we train a GMM on the joint distribution of acoustic feature vectors using the function, `trainGMM`. In the example below, we train two models, one for traditional voice conversion, and another for the accent conversion. The models are saved as Matlab data files so that they can be used during the conversion process.

```
% Traditional VC –pairings based on DTW aligned utterances
[mix,options, errlog]= trainGMM(L1MFCC_FA, L2MFCC_FA);
save gmmModel_crossSpeakerSpectralL1toL2_timealigned.mat mix options
errlog

% AC -pairings based on acoustic similarity
[mix,options, errlog]= trainGMM(L1MFCC_AcSim, L2MFCC_AcSim);
save gmmModel_crossSpeakerSpectralL1toL2_acPair.mat mix options
errlog
```

Step 6: Generate accent modified speech

Given the trained GMM, and a reference test utterance from L1, we generate speech signal that has the linguistic gestures of the test utterance (L1), but the voice-quality of L2. The script below shows how we extract the sequence of L1 acoustic feature vectors `(test_MFCC)` for a test utterance and calculate the sequence of equivalent L2 acoustic features.

```
% load the acoustic features for a given test L1 utterance
testUtt=load('C:\acsouticMappingMethod\L1\mab_a0_0221_STRAIGHT.mat');
test_MFCC = testUtt.MFCC(:,[2:25, 27:50]);
nonSilentFrames = find(~isnan(testUtt.uniquePhLblID));


% Load the GMM model, pitch modification parameters, and the GVs
load gmmModel_crossSpeakerSpectralL1toL2_acPair.mat;
load pitchXForm_L1toL2.mat
load L2TrainUttGVs.mat % GVs from L2 training utterances

% Generate equivalent L2 acoustic features.
% Two estimation methods are available.
% First option estimates acoustic features ignoring their dynamics,
% also known as minimum mean square error criteria
[targetMFCCs_MMSE] =spectralMapping_MMSE(test_MFCC,mix)
wavform = genWavform(testUtt, targetMFCCs_MMSE, pitchXForm);

% The second option estimates the maximum likelihood trajectory
% considering the dynamics and the GVs of the estimated
% acoustic features.
[targetMFCCs_GV_EM] = spectralMapping_MLTrajGV(test_MFCC, mix,
                              trUttGVs , nonSilentFrames);
wavform = genWavform(testUtt, targetMFCCs_GV_EM, pitchXForm);
```

Articulatory-based foreign accent conversion

In articulatory-based accent conversion, we first build an articulatory synthesizer for the L2 speaker, then, drive the synthesizer with the articulatory data from a reference native speaker. StatFAC supports two types of articulatory synthesizer, a GMM-based and a DNN-based. In this example we perform accent conversion using both synthesis models using the articulatory-acoustic corpus available in ConFAC. We generate speech for RGO (the L2 speaker in the corpus) with linguistic gestures (accents) from the reference utterance from MAB (the L1 speaker in the corpus). For the purpose of illustrating StatFAC functionalities, we have broken down the process for articulatory-based accent conversion in six main steps. In the first step, we load the corpus, extract the features, and prepare the training dataset. In the second step, we calculate the pitch parameters for L1 and L2 speakers. In the third step, we calculate the GVs of the training utterances for the L2 speaker. In the fourth step, we train the forward mappings for L2 articulatory synthesizers. We present training method for both GMM-based and DNN-based forward mappings. In the fifth step, we train the articulatory mappings from L1 to L2. Finally, we generate speech with linguistic gestures (accent and style) of a given L1 test utterance, but the voice of the L2 speaker.

Step 1: Prepare the training dataset

In the following, we describe the data preparation script provided with the toolbox. The script (i) loads the corpus provided with ConFAC, (ii) extracts articulatory and acoustic features for all the utterances, and (iii) generates a training dataset for the subsequent processes. The corpus consists of (i) STRAIGHT parameters, (ii) phonetic transcription with timing, and (iii) the drift-corrected trajectories (x-y coordinates in the midsagittal cross-section of the vocal tract) of the EMA pellets for all the utterances. Once the corpus is loaded using ConFAC tools, we create a training dataset from the given set of training sentences. A sample script *(exampleScripts/dataPreparationArtMethod.m)* is provided with the toolbox to create the training dataset for RGO and MAB.

The script extracts acoustic feature vectors ($MFCC_1$ to $MFCC_{24}$ and their deltas) and the articulatory feature vectors (EMA pellet coordinates, $log(f_0)$, *frame energy*, and *nasality*) for all the non-silent frames in the training utterances. For each speaker, the script saves all the features extracted from the training utterances in a Matlab data file. Specifically, the sample script saves the training data for RGO and MAB in the Matlab datafiles named `trDataset_RGO_MFCC_EMA.mat` and `trDataset_MAB_MFCC_EMA.mat`, respectively.

After extracting the features for the training utterances from both the speaker, the script also calculates the phonetic centroids of the EMA pellet positions for both the speakers in the corpus. These phonetic landmarks `(phMeanMAB and phMeanRGO)` are required to train the cross-speaker articulatory mappings. In the case of DNN-based synthesizer, we need the contextualized articulatory input features generated by passing the sequence of articulatory feature vector through a tapped delay line. For this purpose, the script generates two Matlab variables `train_x_multFrames` and `train_y_multFrames` consisting of contextualized articulatory feature vectors and the corresponding acoustic feature vectors, respectively.

Step 2: Calculate parameters for pitch-modification

For each speaker, we calculate the mean and standard deviations of $logf_0$ (log of fundamental frequency) using all the voiced frames in the training dataset. These parameters (stored in a Matlab variable `pitchXForm`) are required to convert pitch trajectory in a reference native utterance to match the range of the non-native speaker.

```
rgoTrData = load('trDataset_RGO_MFCC_EMA');
mabTrData = load('trDataset_MAB_MFCC_EMA');

% pitch transformation from MAB to RGO pitch range
isInputPitch=0; % input is not pitch
[pitchXForm] = calculatePitchTransform(mabTrData, rgoTrData,
isInputPitch)
save pitchXFormMAB2RGO pitchXForm;
```

Step 3: Extract Global Variance (GV)

We now extract the GVs for the L2 speaker (RGO) using only the training utterances. We save the extracted GVs for the conversion step. The example script is given below.

```
% Load the training data for RGO
rgoTrData = load('trDataset_RGO_MFCC_EMA');
[trUttGVs] = calculateGVs(rgoTrData);
save trUttGVsRGO.mat trUttGVs
```

Step 4: Train the forward mappings for L2 synthesizer

GMM-based forward mappings

To train the GMM-based forward mapping for the L2 speaker (RGO in our case), first, we load the training set consisting of the articulatory and acoustic feature vectors. Then, we use the function named `trainGMM` to model the joint distribution of articulatory and acoustic feature vectors using GMMs.

```
rgoTRData = load('trDataset_RGO_MFCC_EMA');
[mix, options, errlog] = trainGMM(rgoTRData.tr_Art,
rgoTRData.tr_MFCC);
save gmmFWDMapRGO.mat mix options errlog
```

DNN-based forward mappings

A sample script to train DNN-based forward mappings for the L2 speaker is given below. Given the articulatory feature vectors(`train_x_multFrames`) and the corresponding acoustic feature vectors (`train_y_multFrames`) contextualized using a tapped-delay line, function `trainDNN` trains the DNN model and returns the model parameters (`M and D`). Model parameters along with the size of the tapped-delay line used to generate training set are saved for future use.

```
Load dnnTrainingData.mat % load the input and output feature vectors
[M,D, zScoreNormParams] = trainDNN(train_x_multFrames,
train_y_multFrames);
save dnnFwdMapRGO.mat M D zScoreNormParams nOfFrames ;
```

Step 5: Train the cross-speaker articulatory mappings

The phonetic centroids of the EMA pellet positions for RGO and MAB are used to train the articulatory mapping function. With these phonetic centroids as the landmarks, we train six Procrustes transforms, one for each EMA pellet. The transforms are later used to register the trajectories of EMA pellet positions in the reference MAB utterances into RGO's articulatory space. The sample script shows the Matlab commands involved in the process.

```
xSpkArtTransforms = trainArtMapping(phMeanMAB, phMeanRGO );
% train L1 EMA ->L2 EMA transforms
save xSpkArtTransforms_mab2rgo.mat xSpkArtTransforms
```

Step 6: Generate accent modified speech

During the conversion stage, we load a reference test utterance from MAB, the trained articulatory mapping transforms, and the pitch modification parameters.

```
% load ConFAC utterance for a test sentence from L1 speaker (say
utt5)
load mabdbUtt.mat
test_utt_id = 5; % the test utterance not used in training
test_mab_u = copyobj(utt_24_all(test_utt_id));
test_mab_u.spk.who = 'mab';
test_mab_u.spk.mainDir='C:\databases\mab_ema\mat\'; % the directory
where the STRAIGHT extracted features for the speaker ‘mab’ in ConFAC
are stored.

load xSpkArtTransforms_mab2rgo.mat; % loads xSpkArtTransforms
load pitchXFormMAB2RGO.mat; % loads pitchXForm
```

Since the conversion process for the GMM and DNN-based approach are different, we describe them separately.

GMM-based approach

In the GMM-based approach, we use `acGMM` function to generate the accent conversion. Given a test utterance from MAB `(test_mab_u)`, the function generates corresponding speech signal in RGO’s voice. The function uses `trajOption` parameter to select one of the three estimation technique supported in StatFAC —the three possible options are explained in the example below.

```
% Trajectory optimization Options
trajOption = 3; % 1: minimum mean square error estimation,
             %      i.e., frame-by-frame mapping
                % 2: maximum likelihood considering dynamics,
                % 3: maximum likelihood trajectory considering the
                % dynamics and GVs of estimated acoustic features

load gmmFWDMapRGO.mat % load the forward-mapping model
[wavform, MFCCs] = acGMM(test_mab_u,xSpkArtTransforms, mix,
                            pitchXForm,trajOption, trUttGVs );
```

DNN-based approach

In the DNN-based approach, we use `acDNN` function to perform accent conversion. As shown in the example given below, the function is used to generate speech signal in the voice of RGO but the linguistic gestures of the given reference test utterance from MAB, `test_mab_u`.

```
load dnnFwdMapRGO.mat;  % M, D, zScoreNormParams and nOfFrames
mu_gv = mean(trUttGVs); % mean GV from training utterances
[wavform, MFCCs] = acDNN(test_mab_u,xSpkArtTransforms, nOfFrames,
                        M,D,zScoreNormParams, pitchXForm, mu_gv);
```