

**ARTICULATORY-BASED SPEECH PROCESSING METHODS FOR  
FOREIGN ACCENT CONVERSION**

A Dissertation

by

DANIEL LEE FELPS

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

DOCTOR OF PHILOSOPHY

August 2011

Major Subject: Computer Engineering

Articulatory-based Speech Processing Methods for Foreign Accent Conversion

Copyright 2011 Daniel Lee Felps

**ARTICULATORY-BASED SPEECH PROCESSING METHODS FOR  
FOREIGN ACCENT CONVERSION**

A Dissertation

by

DANIEL LEE FELPS

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

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,  
Committee Members,

Head of Department,

Ricardo Gutierrez-Osuna  
J. Lawrence Mitchell  
Tracy Hammond  
Jim Ji  
Valerie E. Taylor

August 2011

Major Subject: Computer Engineering

## ABSTRACT

Articulatory-based Speech Processing Methods for Foreign Accent Conversion.

(August 2011)

Daniel Lee Felps, B.S., Texas A&M University

Chair of Advisory Committee: Dr. Ricardo Gutierrez-Osuna

The objective of this dissertation is to develop speech processing methods that enable the modification of a speaker's accent without altering their identity. We envision accent conversion primarily as a tool for pronunciation training, allowing non-native speakers to hear their native-accented selves. With this application in mind, we present two methods of accent conversion. The first assumes that the voice quality/identity of speech resides in the glottal excitation, while the linguistic content is contained in the vocal tract transfer function. Accent conversion is achieved by convolving the glottal excitation of a non-native speaker with the vocal tract transfer function of a native speaker. The result is perceived as 60% less accented, but it is no longer identified as the same individual. The second method of accent conversion selects segments of speech from a corpus of non-native speech based on their acoustic or articulatory similarity to segments from a native speaker. We predict that articulatory features provide a more speaker-independent representation of speech and are therefore better gauges of linguistic similarity across speakers. To test this hypothesis, we collected a custom database containing simultaneous recordings of speech and the positions of important articulators (e.g. lips, jaw, tongue) for a native and non-native speaker. Resequencing speech from a non-native speaker based on articulatory similarity with a native speaker achieved a 20% reduction in accent. The approach is particularly appealing for applications in pronunciation training because it modifies

speech in a way that produces realistically achievable changes in accent (i.e., since the technique uses sounds already produced by the non-native speaker).

A second contribution of this dissertation is the development of subjective and objective measures to assess the performance of accent conversion systems. This is a difficult problem because, in most cases, no ground truth exists. Subjective evaluation is further complicated by the interconnected relationship between accent and identity, but modifications of the stimuli (i.e. reverse speech and voice disguises) allow the two components to be separated. Algorithms to measure objectively accent, quality, and identity are shown to correlate well with their subjective counterparts.

## **DEDICATION**

I dedicate this achievement to Mom and Dad, for their unceasing support and encouragement.

## ACKNOWLEDGEMENTS

I thank my advisor, Dr. Ricardo Gutierrez-Osuna, for his help, support, and dedication to my research. Many of the contributions presented in this dissertation originated from mutual discussions and brainstorming sessions. I am grateful for his guidance and enthusiasm for research. Above all else, he challenged me to become a better researcher.

I thank my committee members, Dr. Larry Mitchell, Dr. Heather Bortfeld, Dr. Jim Ji, and Dr. Veronica Loureiro-Rodriguez for their interest in my work. Christian Geng played a vital role by recording the articulatory database that enabled this research. I would also like to thank Michael Berger and Ricardo Gutierrez-Osuna for serving as the native and non-native speakers. I am grateful to Professor Hideki Kawahara for making his Straight speech processing algorithm available to me and for the financial support provided by the Science, Mathematics and Research for Transformation Scholarship program from the Department of Defense.

I will miss past and present members of the Pattern Recognition and Intelligent Sensor Machines (PRISM) Lab, who have been excellent colleagues and friends over the last 6 years. I would like to especially thank Jobany, Jong, and Rakesh for keeping my spirits up when “life was tough.” Finally, I would like to thank my family for their unconditional love and support.

## TABLE OF CONTENTS

	Page
ABSTRACT .....	iii
DEDICATION .....	v
ACKNOWLEDGEMENTS .....	vi
TABLE OF CONTENTS .....	vii
LIST OF FIGURES .....	x
LIST OF TABLES .....	xiii
1. INTRODUCTION.....	1
1.1 Organization of this document .....	2
2. BACKGROUND.....	4
2.1 Overview of linguistic terms .....	4
2.2 Accent .....	5
2.2.1 Native speaker reactions to non-native speech .....	5
2.2.2 Relationship between accent and identity .....	8
2.2.3 Spanish-English specific differences .....	9
2.3 Speech production.....	12
2.3.1 Acoustic theory of speech production.....	13
2.3.2 Modulation theory of speech production .....	15
2.3.3 Front cavity hypothesis.....	16
2.3.4 Language specific articulatory settings.....	17
2.3.5 Articulatory corpora.....	19
2.4 Synthesis and modification of speech .....	22
2.4.1 Speech synthesis .....	22
2.4.2 Speech modification .....	26
2.4.3 Voice conversion .....	27
3. LITERATURE REVIEW .....	31
3.1 Accent conversion.....	31
3.2 Speech modification applications in pronunciation training.....	35
3.3 Quantifying accent .....	36



	Page
3.3.1 Automatic classification of accents .....	37
3.4 Quantifying quality .....	38
3.4.1 Automatic assessment of quality .....	39
3.5 Quantifying identity .....	40
3.5.1 Automatic classification of speaker identity .....	41
4. SPECTRAL FOREIGN ACCENT CONVERSION .....	43
4.1 Speech modification framework .....	43
4.2 Accent conversion .....	48
4.2.1 Prosodic conversion .....	48
4.2.2 Segmental conversion .....	49
4.3 Experimental .....	52
4.3.1 Results .....	54
4.3.2 Discussion .....	57
4.4 Conclusion .....	59
5. OBJECTIVE MEASURES OF ACCENT CONVERSION .....	60
5.1 Acoustic quality .....	60
5.2 Foreign accent .....	62
5.3 Speaker identity .....	64
5.4 Experiment .....	66
5.4.1 Results .....	66
5.4.2 Discussion .....	69
5.5 Conclusion .....	72
6. CONCATENATIVE FOREIGN ACCENT CONVERSION .....	73
6.1 ConFAC analysis .....	76
6.2 Accent conversion .....	77
6.2.1 Feature normalization .....	79
6.3 ConFAC synthesis .....	80
6.3.1 Concatenative unit selection .....	81
6.3.2 Optimal coupling .....	85
6.3.3 Spectral smoothing with pulse-density modulation .....	87
6.3.4 STRAIGHT synthesis .....	89
6.4 Conclusion .....	91
7. CONFAC EXPERIMENTS .....	92
7.1 Subject recruitment .....	92
7.2 Experiment #1 – Accent rating .....	93
7.2.1 Results .....	94
7.3 Experiment #2- Decoupling accent and identity (Part 1) .....	95
7.3.1 Results .....	97

	Page
7.4 Experiment #2- Decoupling accent and identity (Part 2).....	97
7.4.1 Results.....	98
7.5 Experiment #3 - Comparing features for cross-speaker synthesis .....	99
7.5.1 Results.....	101
7.6 Discussion .....	103
8. CONCLUSION .....	106
8.1 Future work.....	109
REFERENCES .....	112
APPENDIX A .....	123
APPENDIX B.....	127
APPENDIX C.....	141
APPENDIX D .....	143
APPENDIX E.....	145
APPENDIX F .....	149
VITA .....	165

## LIST OF FIGURES

	Page
Figure 1 A comparison of Spanish and American English vowel systems. ....	11
Figure 2 Midsagittal diagram of human speech production mechanisms. ....	12
Figure 3 According to the acoustic theory of speech production, speech is the convolution of a source (left) with a filter (middle).....	14
Figure 4 Wideband spectrogram of the sentence, “How much was it?” .....	14
Figure 5 General discrete-time model of speech production (Deller et al., 2000). ....	14
Figure 6 Tracings of mid-sagittal x-rays of the vowels /i/ and /a/ produced by an adult male and a child. ....	17
Figure 7 Articulatory trajectories for the phrase “prepared for deployment.”.....	21
Figure 8 Articulatory speech synthesis mimics the human speech production process. ....	24
Figure 9 Klatt’s hybrid serial/parallel formant synthesizer uses 19 parameters to synthesize speech [reprinted from (Schroeter, 2008)]. ....	24
Figure 10 PSOLA modification of speech. ....	28
Figure 11 STRAIGHT parameterization of speech consists of (a) smooth spectrogram and (b) aperiodicity signal ( fundamental frequency not shown).....	28
Figure 12 A typical voice conversion system. A mapping is learned from the source to the target speaker in the training phase. ....	30
Figure 13 Comparison of two spoken language conversion systems to synthesize English utterances from a corpus of Japanese speech.....	33
Figure 14 A summary of the FD-PSOLA framework. ....	45
Figure 15 Pitch lowering in the frequency domain. ....	47
Figure 16 Creating the PSOLA synthesis pitch marks. ....	47

	Page
Figure 17 Modifying the FD-PSOLA framework (Figure 14) for the segmental conversion.....	50
Figure 18 VTLN frequency mapping is created by linearly interpolating between average formant locations for the foreign and native speakers.....	51
Figure 19 Wideband spectrograms of the utterance “...and her eyes grew soft and moist.”.....	51
Figure 20 Accent and quality ratings for SpFAC.....	55
Figure 21 Experimental results from the identity tests.....	57
Figure 22 Calibrating the HTK score.....	64
Figure 23 (a) Correlation between objective and subjective measures of accent.....	68
Figure 24 (a) Correlation between objective and subjective measures of acoustic quality.....	68
Figure 25 (a) Correlation between objective and subjective measures of identity.....	69
Figure 26 Average location of 11 vowels in a low dimensional (a) acoustic and (b) articulatory vowel spaces.....	74
Figure 27 ConFAC system overview for articulatory-driven accent conversion.....	75
Figure 28 ConFAC’s analysis stage encodes an utterance into a format that permits unit selection based synthesis.....	76
Figure 29 ConFAC’s accent conversion module compares the native and non-native synthesis features.....	77
Figure 30 Creating the mispronunciation label file for the word “anything” to identify differences between the non-native and native phonetic sequences.....	78
Figure 31 Calculating the Maeda parameters from EMA (x,y) positions.....	80
Figure 32 ConFAC’s synthesis stage creates an acoustic waveform from the synthesis features.....	81
Figure 33 Representing a speech corpus as a state transition network.....	83
Figure 34 The left diphone (a) and right diphone (b) are joined to form a triphone (c).....	86

	Page
Figure 35 Calculating the cost of joining two diphones. ....	87
Figure 36 Morphing two spectrums with the PDM method. ....	88
Figure 37 STRAIGHT parameterization of the MAB utterance “Zimbabwe is the first example.” .....	90
Figure 38 Accent ratings for ConFAC. ....	94
Figure 39 Transforming a speaker to a particular guise. ....	96
Figure 40 Accent ratings for the change of identity experiment. ....	97
Figure 41 Accent ratings for the experimental conditions after undergoing a change of identity.....	99
Figure 42 Graphical representation of the relationship between accent and identity. ....	109
Figure 43 Calculating a complete set of geodesic distances from a subset of utterance pairings.....	142
Figure 44 Long term drift correction.....	144
Figure 45 The pre-qualification test. ....	146
Figure 46 The accent rating form used in ConFAC experiments #1 and #2 (sub- sections 7.2-7.4).....	147
Figure 47 The evaluation form for the third ConFAC experiment.....	148
Figure 48 Diagram showing the relationship between the Matlab classes.....	152
Figure 49 Basic interaction with an utterance object.....	155
Figure 50 A script demonstrating the standard way to load data and prepare it for accent conversion.....	156
Figure 51 A simple text-to-speech example using ConFAC.....	160
Figure 52 Example code for performing accent conversion with ConFAC.....	161
Figure 53 Adding a different speaker type to ConFAC.....	162
Figure 54 Modifying the utterance class to add a new feature.....	164

## LIST OF TABLES

	Page
Table 1 The experimental conditions of a voice identification task.....	9
Table 2 Types of information and variation in speech [reprinted from (Traunmüller, 1998)]. .....	16
Table 3 Relative differences in articulatory settings for British-English and French languages (Honikman, 1964). .....	19
Table 4 Stimulus conditions for perceptual studies.....	52
Table 5 Combined score for identity ratings. ....	53
Table 6 Transcripts of the 10 sentences used in Experiments 1 and 2 of this section. ....	93
Table 7 Six conditions used in part 1 of Experiment #2.....	96
Table 8 Separation of stimuli into two test sets (A and B). ....	98
Table 9 Experimental conditions in the cross-speaker synthesis test. ....	100
Table 10 Contingency table showing the results of the pre-test. ....	102
Table 11 The properties of the “utterance” class are listed below.....	153
Table 12 Various parameters and their effect on ConFAC synthesis.....	159

## 1. INTRODUCTION

Despite years or decades of immersion in a new culture, older learners of a second language (L2) typically speak with a so-called “foreign accent,” sometimes despite concerted efforts at improving pronunciation. Among the many aspects of proficiency in a second language (e.g., lexical, syntactic, semantic, phonological), native-like pronunciation can be the most difficult to master because of the neuro-musculatory basis of speech production (Scovel, 1988). A foreign accent does not necessarily affect a person’s ability to be understood (Munro and Derwing, 1995), but it may subject them to discriminatory attitudes and negative stereotypes (Anisfeld et al., 1962; Arthur et al., 1974; Lippi-Green, 1997; Ryan and Carranza, 1975; Schairer, 1992). Thus, by achieving near-native pronunciation, L2 learners stand to gain more than just better intelligibility.

During the last two decades, a handful of studies have suggested that it would be beneficial for L2 speakers to be able to listen to their own voices producing native-accented utterances (Jilka and Möhler, 1998; Sundström, 1998; Tang et al., 2001; Watson and Kewley-Port, 1989). The rationale is that, by stripping away information that is only related to the teacher’s voice quality, it is easier to perceive differences between their accented utterances and accent-free counterparts. This dissertation seeks to transform foreign-accented speech into its native-accented counterpart. The problem of accent conversion (AC) is related to but distinct from that of voice-conversion (Stylianou et al., 1998). Whereas voice conversion seeks to transform the voice of a speaker to sound like a different speaker, accent conversion seeks to

---

This dissertation follows the style of Speech Communication.

transform only those features of an utterance that contribute to accent while maintaining those that carry the identity of the speaker.

This dissertation investigates several issues surrounding the automatic generation of accent-modified speech. Two methods of AC are presented: 1) spectral foreign accent conversion (SpFAC), which modifies non-native speech based on the source/filter decomposition of speech, and 2) concatenative foreign accent conversion (ConFAC), which resequences existing non-native speech based on acoustic or articulatory similarity to a native speaker. The primary objective of the work is to explore the potential benefits of relying on physical movements of the tongue, lips, and jaw to perform accent conversion. We hypothesize that the articulatory domain is better suited to separate features that cue accent from those related to voice quality. To achieve this objective we collected two specialized corpora (one from a native of American English; another from a non-native speaker) containing simultaneous recordings of articulatory and acoustic waveforms. Accent conversion was then performed in both domains (articulatory and acoustic) and evaluated using subjective and objective measures designed specifically for this work.

### **1.1 Organization of this document**

This dissertation is organized as follows. Section 2 reviews perceptual and social experiments involving foreign accented speakers as well as selected speech processing methods that are used throughout the dissertation work. Section 3 provides a literature review of related research and previous approaches to accent conversion. Section 4 presents an accent conversion system based on the source/filter decomposition of speech (SpFAC) and evaluates its effect on the perception of accent, quality, and identity. Section 5 proposes ways to measure these criteria objectively. Section 6 presents an accent conversion system based on concatenative speech synthesis (ConFAC). Section 7 evaluates ConFAC and compares the use of acoustic or



articulatory features for accent conversion. The final section presents directions for future research.

## 2. BACKGROUND

This section reviews various topics concerning accent, including the perception of foreign-accented speech among native speakers, the automatic detection of accent, and the use of computer tools to improve one's pronunciation of a particular language. It also discusses general theories of speech production to motivate our choice of using articulatory (rather than acoustic) features to perform accent conversion. Finally, it reviews select speech processing algorithms relevant to this dissertation.

### 2.1 Overview of linguistic terms

Among the linguistic terminologies found throughout the manuscript, the most fundamental concept is that of the *phoneme*. A phoneme is the smallest structural unit that distinguishes meaning in a language. For example, the word “phoneme” is phonetically spelled using the six phonemes /foonim/ (see APPENDIX A for a full listing). The related term *phone* is used to describe a particular instance of a phoneme in a real utterance. However speech is much more than a sequence of phones; the actual production of a phone is influenced by several factors including the surrounding phones (coarticulation), the intent of the message (e.g. is the speaker informing, requesting, apologizing, or disagreeing?), and speaker-dependent factors (e.g. accent and emotion). Phones vary by stress (energy of a sound), length (duration of a sound), tone (short pitch variation, i.e. that which changes the meaning of a sound in tonal languages), and intonation (long pitch variation, i.e. the difference between a question and a statement). The collection of these descriptors makes up the prosody of speech. Prosody helps listeners parse speech and it also conveys information related to syntax and pragmatics.

## **2.2 Accent**

The term “accent” describes the way a speaker produces the sounds of a language. An accent can indicate the speaker’s first language (native or non-native), where they were born (regional accent), religious affiliation, ethnic group, or socio-economic class. Accents affect both acoustic (e.g. formants) and prosodic (e.g. intonation, duration, and rate) aspects of speech. The related term “dialect” describes a person’s accent in addition to their vocabulary and grammar. However, the proposed speech processing algorithms modify only the segmental and prosodic aspects of speech and are therefore referred to as methods of “accent conversion.”

### **2.2.1 Native speaker reactions to non-native speech**

Various research paradigms have been used to measure the effect of accent on listener evaluations of personality, intelligence, socioeconomic status, and degree of desired interaction with the speaker (Arthur et al., 1974; Brennan and Brennan, 1981; De La Zerda and Hopper, 1979; Fielding and Evered, 1980; Giles and Powesland, 1975; Kalin and Rayko, 1978; Ryan et al., 1977; Ryan and Carranza, 1975; Strongman and Woosley, 1967). The most powerful of these paradigms is the matched guise experiment first proposed by (Lambert et al., 1960), which uses bilingual speakers to control for other variables such as voice quality and speaker personality. Such speakers typically learn both languages as a child, which is common in situations where each parent speaks a different native language or the household language is different than the schooling language. Each accent then becomes a “guise” for that speaker; naïve listeners will mistakenly assume it is two different people.

Strongman and Woosley (1967) conducted a matched guise experiment with bilingual speakers who could effortlessly switch between Yorkshire and London accents (Yorkshire is located in northern England and London is located in southern England). Evaluative reactions from northern and southern English participants were collected. The authors found that

participants from either location rated the London guises to have higher speaker confidence and the Yorkshire guises to have higher honesty, reliability, and generosity. Another significant finding was that only northern judges rated the Yorkshire guises to be less irritable and more good natured, kind hearted, and industrious; Giles and Powesland (1975) later termed this kind of one-sided finding as “accent loyalty.”

Accent discrimination does not only stem from geographic differences. Anisfeld et al. (1962) performed a matched guise using Jewish speakers who also had the ability to speak with a standard Canadian-English accent. Both Jewish and non-Jewish participants rated the guises on 14 traits, which spanned from physical evaluations (e.g. height or good looks) to personal evaluations (e.g. humor or kindness). Both groups rated the Canadian-English guises more favorably on height, good looks, and leadership. In another example of accent loyalty, the Jewish participants rated the Jewish guises more favorably on sense of humor and kindness.

A few authors have investigated the effect of different degrees of the same accent on evaluations of socioeconomic status. Brennan and Brennan (1981) recorded nine Mexican Americans with varying degrees of accent. The speakers were then evaluated by a panel of linguists to assign each speaker an accent index based on 18 pronunciation variables. The speakers were evaluated by 43 Mexican American and 37 Anglo American high school students; each student rated the speakers on *status variables* (e.g. level of education or success) and *solidarity variables* (e.g. trustworthiness or friendliness). They found accent index to be highly correlated with the status variables (i.e. the higher the accent, the lower the status). These results parallel those found independently by Ryan et al. (1977), who found that small increases in accentedness were associated with gradually less favorable ratings of status and solidarity. Furthermore, Sebastian et al (1979) concluded that not only were Spanish accented speakers

thought to be in a lower social class, but the participants also had less desire to be in a social relationship with them.

These previous studies tested the ability of accent to convey stereotypes and prejudices in an academic setting (i.e. through a questionnaire). Other studies have shown that non-native speakers can be at a serious disadvantage in real-world situations such as when looking for employment and housing. One investigation (De La Zerda and Hopper, 1979) asked employers to evaluate simulated interviews with potential employees with various degrees of Mexican-American accent. The employers rated each potential employee with respect to three job positions: supervisor, skilled technician, and semi-skilled worker. The results show that employers favored standard English speakers as supervisors and accented speakers for the semi-skilled worker position. In a similar experiment, Kalin and Rayko (1978) tested if this effect was second-language dependent. For this purpose, the authors asked Canadian employers to rate four varieties of speakers for four levels of job status. Employers favored Canadian-English speakers for the highest status jobs over German, then south-Asian, and finally west-Indian speakers; the order was reversed for the lowest status jobs.

Housing is a scenario where potential tenants are at the mercy of landlords. Purnell et al. (1999) employed a matched guise experiment with a single talented speaker who was able to speak three dialects of American English: standard, African American vernacular, and Chicano. The speaker called a landlord three times over a short period requesting to look at an apartment. The results show that landlords discriminated against prospective tenants on the basis of their accent. Further analysis revealed that the number of callbacks for African American and Chicano guises was positively correlated with the demographics of the neighborhood (i.e. fewer callbacks in areas with small minority populations). Another study uncovered a surprising relationship between accent and the diagnosis of an illness. Fielding and Evered (1980) asked physicians to

diagnose an illness from a recording of a speaker describing their illness. They found that the physicians were more likely to diagnose a speaker with a Received Pronunciation English accent as having a psychosomatic problem, whereas speakers with a rural English accent were more likely to be diagnosed with a physical problem. In conclusion, even if a speaker's accent does not hinder their ability to be understood, they may still be stereotyped by those with whom they communicate.

### **2.2.2 Relationship between accent and identity**

The main goal of accent conversion (i.e., altering the perceived accent of a speaker) simultaneously and unavoidably alters the perceived identity. Criminals are aware of this fact and, in addition to disguising their face, often disguise their voice; statistics from the German Federal Police Office show that criminals employed some form of voice disguise (e.g. creaky voice, whispering, faking an accent, or pinching one's nose) in 15-25% of criminal cases involving speaker identification (Künzel, 2000). Sjöström et al. (2006) simulated this experience to determine the effect of a dialect switch on voice identification. They employed a bidialectal speaker who was born near Stockholm (ST), but moved to Scania (SC) when he was five years old. As an adult, he continued to use both dialects on a daily basis. Stimuli for a voice line-up were collected from the speaking talent and four foil speakers (2 mono-dialectal speakers from either SC or ST). The experiment was composed of a familiarization stage and an identification stage. Participants first became familiar with one of the dialects from the bidialectal speaker (ST or SC) by listening to a recorded passage. They were then presented with one of the four voice line-ups in Table 1 and asked to select the speaker that most closely resembles the voice used during familiarization. The results show that the speaker was not recognized when he switched dialects (i.e. ST-SC and SC-ST). Thus, dialect is an important clue for speaker identification.

This issue complicates the evaluation of accent conversion and reinforces the ambiguous nature of the result.

Table 1

The experimental conditions of a voice identification task. This study investigated the potential use of accent as a voice disguise [reprinted from (Sjöström et al., 2006)]. Discrimination sensitivity is measured using the  $d'$  sensitivity index from signal detection theory<sup>1</sup>.

Test	Familiarization voice	Line-up voices	$d'$ value
SC-SC	TargetSC	Foil 1-4 + TargetSC	1.87
ST-ST	TargetST	Foil 1-4 + TargetST	1.93
ST-SC	TargetST	Foil 1-4 + TargetSC	0.44
SC-ST	TargetSC	Foil 1-4 + TargetST	-0.07

### 2.2.3 Spanish-English specific differences

Though some aspects of a foreign accent may be specific to the individual, most can be predicted from the characteristics of the first and second language (L1 and L2). This dissertation focuses on Spanish accented English, which is the most prominent foreign accent found in Texas. This sub-section compares the phonetic and prosodic differences unique to Spanish speakers of English, although there are many other differences beyond the scope of this dissertation (e.g. orthographic, punctuation, grammar, and vocabulary).

The English vowel system, with 12 monophthongs, is especially difficult for Spanish speakers since the Spanish vowel system has only 5. Figure 1 identifies differences between the two vowel sets; uniquely English vowels are often pronounced/heard as the nearest Spanish vowel. Namely, *seat* /sIt/ sounds like *sit* /sit/, *caught* /kɔt/ like *coat* /cot/, and *pool* /pul/ like *pull*

<sup>1</sup> Chance level for  $d'$  is 0. The  $d'$  scores for ST-SC and SC-ST conditions are not significantly different than 0,  $t(39)=1.36$ ,  $p>0.05$ , indicating random response (Sjöström et al., 2006).

/pʊl/. The English vowels /æ/, /ɑ/, and /ʌ/ are often mapped to the Spanish vowel /a/, which leads to confusion among *cart*, *cat*, and *cut*. The English central vowel /ə/ is a reduction of primary vowels that appear in unstressed positions, but Spanish speakers often use the primary form instead (e.g. *about* /əbaut/ is pronounced as /about/) (Coe, 2001).

Spanish and English share 17 identical or near equivalent<sup>2</sup> consonants, though this does not prevent such phones from being used incorrectly. For example, Spanish speakers often drop /k/ following /ŋ/ as in *sink* or replace /m/ for /n/ in a final position turning *dream* into *drean*. The most problematic English consonants are fricatives. Namely the English phones /z/, /dʒ/, /tʃ/, /ʃ/, are frequently confused (compare *pleasure*, *plejure*, *pletcher*, *plesher*). Of the four, only /tʃ/ exists in Spanish<sup>3</sup>. The Spanish /s/ is also the closest phone to the English /z/.

Certain English consonant clusters are also problematic. A common example is /s/ appearing with another consonant at the beginning of a word. Since this never occurs in Spanish, Spanish speakers will often add an initial /e/ (e.g. *Espain*). Other examples include *espres* for *express*, *brefas* for *breakfast*, or *win* for *wind*.

---

<sup>2</sup> An example of a “near equivalent” consonant is the English approximant /ɹ/ compared to the Spanish alveolar flap /r/. Although these sounds are not identical, they are difficult to differentiate in continuous speech. See Coe (2001) for a comprehensive list.

<sup>3</sup> The Spanish /s/ can sound like /ʃ/ in certain contexts, e.g. *see* may be pronounced as *she*.



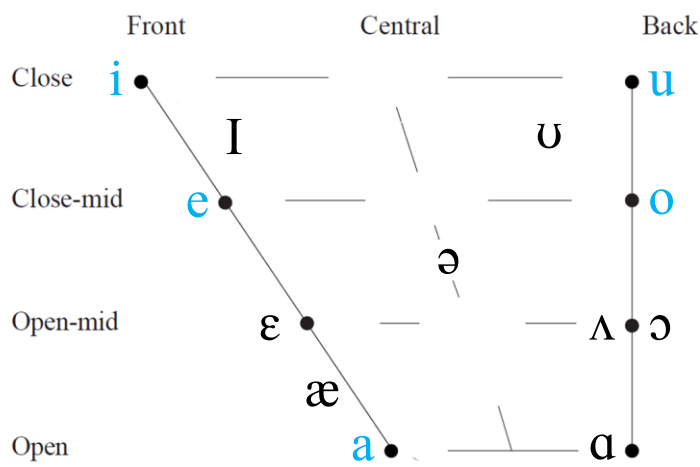


Figure 1 A comparison of Spanish and American English vowel systems. Spanish vowels are shown in blue. The only vowel that is not an English vowel is open-front vowel /a/, which is closest to /æ/ in *fast*.

Spanish and English also have pronunciation differences at the suprasegmental level. Nava et al. (2009) investigated the relationship between rhythm and prominence and its applications to automatic pronunciation scoring for Spanish learners of English. They summarize the differences between English and Spanish prosody as: “*English is considered a “stress-timed” language due to the presence of vowel reduction, varied syllable structure inventory including complex onsets and codas, and vowels in stressed syllables that are regularly longer than in unstressed syllables. Spanish, on the other hand, is considered “syllable-timed” and does not have vowel reduction, has a reduced syllable inventory in comparison with stress-timed languages, and the difference between stressed and unstressed vowels is not as great*” (Nava et al., 2009). The authors analyze a corpus of native and nonnative English speakers to show that phrasal prominence is a good indicator of overall pronunciation ability. In fact, prosody is so important to American English that it has been used to classify three different American accents (i.e. typical accents from California, Mississippi, and New York) (Ikeno and Hansen, 2006).

### 2.3 Speech production

This sub-section reviews the anatomical act of speech production and gives an account of speech production models that motivate the proposed methods of accent conversion. The human speech production system relies on “articulators” (e.g. vocal folds, tongue, teeth, jaw, lips, and velum) to modify the configuration of the vocal tract and produce sound (Figure 2). Speech production begins as air leaves the lungs and passes through the trachea and vocal folds. If the vocal folds are contracted, they vibrate to produce a periodic excitation to the rest of the production system. The frequency of the vibrations is called the fundamental frequency and the perception of this frequency is called pitch. Depending on whether this vibration occurs, sounds are classified as voiced or unvoiced.

Given both the glottal excitation and vocal tract shape, an acoustic waveform is uniquely defined (Fant, 1970; Flanagan, 1972; Schroeter and Sondhi, 1994; Stevens, 1998). This is known as the forward or articulatory-to-acoustic mapping; it is a univalued mapping (although highly nonlinear) that allows speech to be synthesized from articulatory parameters. This complex process is often approximated as a linear system using the acoustic theory of speech production, which we review next.

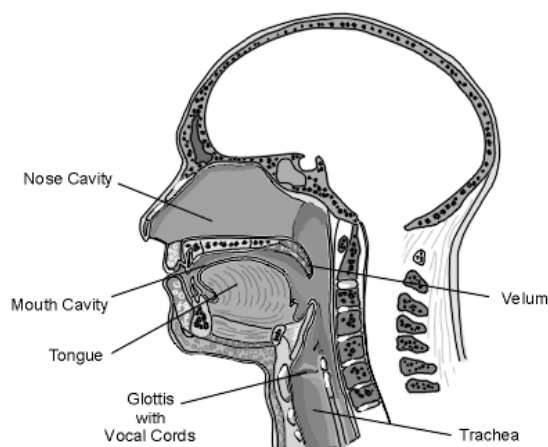


Figure 2 Midsagittal diagram of human speech production mechanisms.

### 2.3.1 Acoustic theory of speech production

Introduced by Fant (1970), the acoustic theory of speech is often referred to as the source-filter theory because it models speech as the interaction between two components: a source and a filter (Figure 3). The source is created as air from the lungs passes through the vocal folds. The filter is defined by the configuration of the oral and nasal tracts. Each configuration resonates at certain frequencies, just as a guitar string's resonant frequencies are determined by its length. Those areas of concentrated energy are known as formants, and they appear as dark bands in spectrograms (see Figure 4).

Source-filter theory is useful due to the fact that it can be implemented as a slowly varying discrete-time linear system. Speech can be considered slowly varying for short periods because it is quasi-periodic (i.e. the speech signal is effectively stationary for periods of about 10-30 ms). During these brief glimpses, the articulators are relatively fixed. Shown in Figure 5, voiced speech begins when an impulse train (whose frequency is determined by the pitch period) is convolved with a glottal filter, which shapes the impulses into a smooth glottal pulse waveform. Unvoiced speech, on the other hand, is produced with a white noise generator. From there, the vocal tract filter shapes the source. A final filter relates the pressure and volume velocity of speech at the lips. The acoustic theory of speech production provides motivation for spectral foreign accent conversion, which will be presented in section 4.

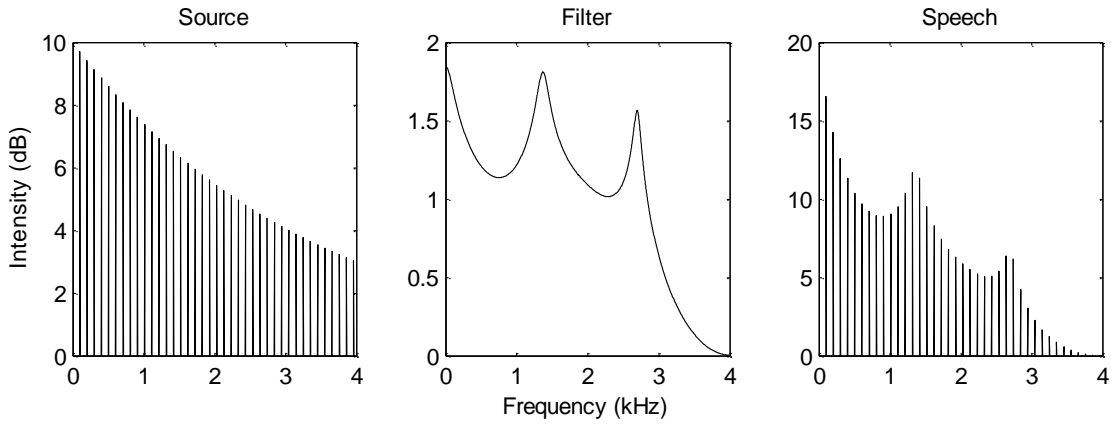


Figure 3 According to the acoustic theory of speech production, speech is the convolution of a source (left) with a filter (middle).

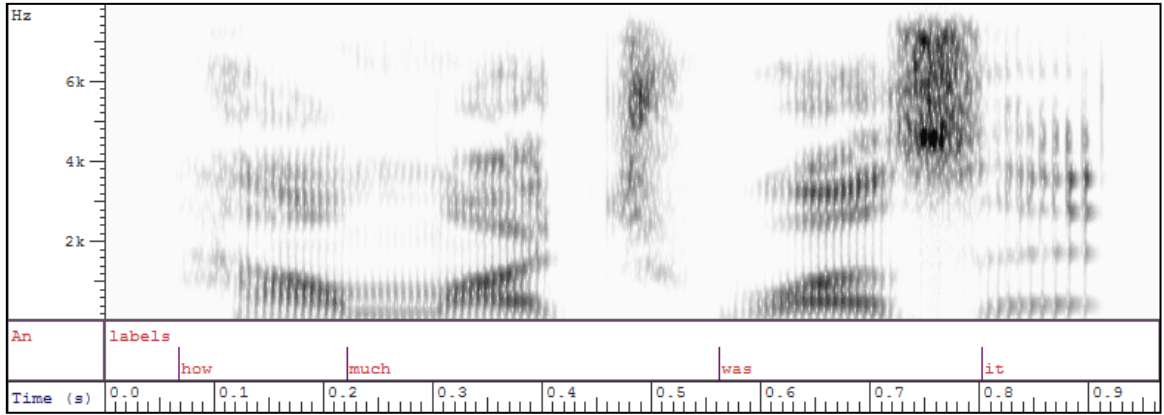


Figure 4 Wideband spectrogram of the sentence, “How much was it?”

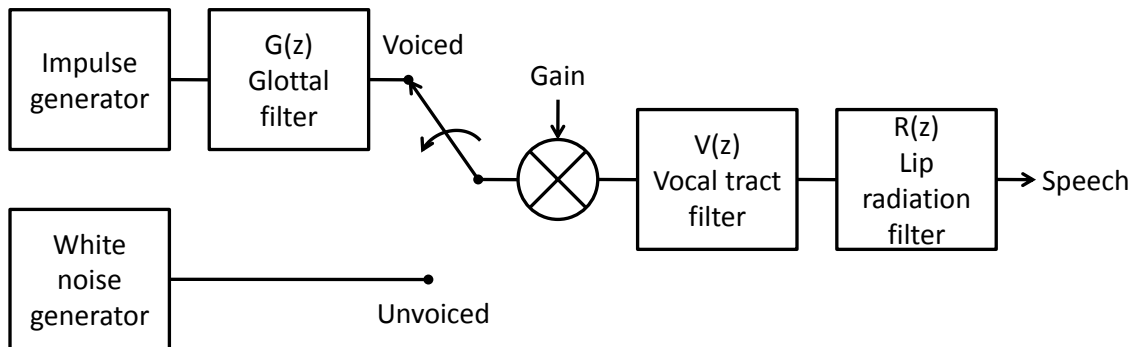


Figure 5 General discrete-time model of speech production (Deller et al., 2000).

### 2.3.2 Modulation theory of speech production

Source-filter theory provides a functional model of speech production that is simple, accurate, and easy to represent mathematically. Modulation theory (Traunmüller, 1994) provides a broader, more abstract view of speech production. Modulation theory regards speech as a combination of linguistic, expressive, organic, and perspectival qualities (Table 2). It characterizes speech as the complex modulation of a carrier signal by articulatory gestures: this carrier signal is “*an unarticulated, ‘colorless’ and linguistically featureless oral vowel, produced with the vocal folds optimally adduced for phonation and relaxed (slack) to the extent to which expressive factors allow*” (Traunmüller, 1998). This neutral sound is determined by the morphological and biomechanical properties of the speech organs. Therefore, according to Modulation theory, the speaker’s characteristics are not solely determined by the glottal source (as in source-filter theory), but also by a neutral vocal tract. Modulation theory also seeks to explain speech perception by viewing it as a *demodulation* process, which allows listeners to recover the message once they “tune into” the carrier of the speaker and the idiosyncratic way in which the speaker modulates the signal (i.e. speaks). According to this view, then, a foreign accent may be removed from an utterance by extracting its voice-quality carrier and convolving it with the linguistic gestures of a native-accented counterpart.

Table 2

Types of information and variation in speech [reprinted from (Traunmüller, 1998)].

Quality and characterization	Information conveyed	Phenomena involved
Linguistic Social, conventional	message, dialect, accent, speech style	Words, speech sounds, prosodic patterns
Expressive Psycho-psychological, within speaker variation	Emotion, attitude, environment	Vocal effort speaking rate, pitch dynamics, voice quality
Organic Morphological, between speaker variation	Age, sex, pathology	Larynx size, vocal tract length
Perspectival Spatial, transmittal	Place, distance, orientation, transmission channel	Acoustic and optic factors

### 2.3.3 Front cavity hypothesis

The specific motivation for using articulatory information to perform accent conversion stems from work by Kuhn (1975) and Hermansky and Broad (1989), which suggests that “*the size and shape of vocal tract’s front cavity is the primary carrier of linguistic information. The back cavity geometry is its causal consequence and contributes mainly speaker-dependent information.*” Thus, the frontal cavity hypothesis (FCH) provides an approach for addressing the key challenge in accent conversion: separating linguistic information from that which is speaker-dependent. Namely, according to this hypothesis, the front cavity of the vocal tract captures the linguistic information. Support for this hypothesis is shown in Figure 6, which depicts the vocal tract (mid-sagittal x-ray tracings) for an adult male and a child for two stationary vowels. As predicted by FCH, both oral cavities have similar shapes, but the back cavity is much larger in the adult. A handful of opposing studies report significant differences between speakers’ articulatory configurations, but none are able to determine if these differences affect production (Hashi et al., 1998; Johnson et al., 1993; Simpson, 2001; Westbury et al., 1998).

Thorough testing of the front cavity hypothesis requires data collection on the entire vocal tract. Magnetic Resonance Imaging (MRI) is currently the best way to obtain such

information (Story et al., 1996), but there are no public, multiple speaker MRI databases. Instead, this dissertation uses partial vocal tract information (going as far back as the tongue dorsum) collected using an electromagnetic articulograph. We test the similarity of articulatory trajectories for two speakers using a custom speech synthesizer that can be driven by articulatory features. This is a significant improvement over previous studies that measure similarly directly in the articulatory domain because it allows us to estimate the acoustic consequences of articulatory differences.

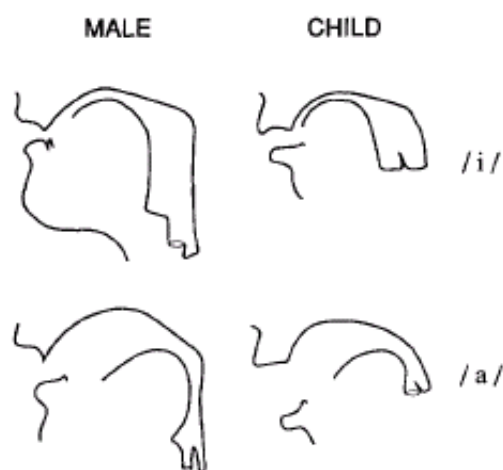


Figure 6 Tracings of mid-sagittal x-rays of the vowels /i/ and /a/ produced by an adult male and a child. Though the adult's vocal tract will be much longer than the child's, the front cavity is approximately the same shape. This suggests that the front of the vocal tract captures the linguistic information while the back carries mostly speaker dependent information [reprinted from (Hermansky and Broad, 1989)].

### 2.3.4 Language specific articulatory settings

Further motivation for articulatory-based accent conversion can be found in studies of language-dependent articulatory settings. Honikman (1964) defines articulatory settings as the “*gross oral posture and mechanics required for the economic and fluent production of the established pronunciation of a language.*” Language-dependent articulatory differences are primarily a matter of efficiency since languages have different distributions of sounds. For

example, Indian languages are spoken with the jaws held relatively loosely, facilitating the production of frequently occurring retroflex consonants. Relative differences between French and British-English articulatory settings are summarized in Table 3 (Honikman, 1964).

The existence of language-dependent articulatory settings has important implications for second language learning. It suggests that non-native speakers of a language can only achieve fluent pronunciation of that language by learning its default articulatory settings (in addition to its phonemes and prosody). Furthermore, a speaker's foreign accent may be partially explained by using a foreign articulatory setting. If this is this case, then accent conversion may benefit from articulatory information. Admittedly, the articulatory data used in this dissertation is not complete enough to detect most of the differences listed in Table 3, but it may contain indirect evidence of such occurrences. As an example, speakers of British-English generally anchor their tongue laterally to the roof of the mouth, but our dataset cannot detect this because it only measures the vocal tract at points along the midsagittal plane. However, lateral anchoring of the tongue restricts its entire movement including those along the midsagittal plane. We next discuss the articulatory dataset in more detail.



Table 3  
Relative differences in articulatory settings for British-English and French languages (Honikman, 1964).

<i>Setting</i>	<i>British-English</i>	<i>French</i>
Jaw	loosely closed	slightly open
Lip	neutral; moderately active	rounded; vigorously active
Oral cavity	relaxed	cheeks contracted
Main consonant articulator	tip – alveolar	blade – dental
Tongue anchorage	to roof laterally	to floor centrally
Tongue tip	tapered	untapered
Tongue body	slightly concave to roof	convex to roof
Tongue underside	concave to roof	neutral

### 2.3.5 Articulatory corpora

A special database was collected for this research at the Centre for Speech Technology Research, University of Edinburgh in Fall 2009 using a Carstens AG500 3D-articulograph. The non-native subject (RGO) is from Madrid, Spain. English is his second language and he began studying it at age 6, but he primarily spoke Spanish until he moved to the United States at the age of 25. At the time of the recording he was 41 years old and had been living in the U.S. for 16 years. The native speaker (MAB) is a monolingual speaker who grew up in New York; he was 39 years old at the time of the recording. Both subjects recorded the same 344 sentences chosen from the Glasgow Herald corpus (APPENDIX B). In addition, RGO recorded 206 sentences not spoken by MAB. Audio recordings were captured at a sampling rate of 32 kHz with an AKG CK98 shotgun microphone. Articulatory features were simultaneously recorded through the use of ten electromagnetic pellets— four were used to cancel head motion and provide a frame of reference, while the other six were attached to capture articulatory movements (upper lip, lower lip, jaw, tongue tip, tongue mid, and tongue back); the front-most tongue sensor was positioned 1cm behind the actual tongue tip, the rearmost sensor (TD) as far back as possible without creating discomfort for the participant, and the third sensor was placed equidistant from TT and

TD (Hoole et al., 2003). Example trajectories for the phrase “*Prepared for deployment*” are shown in Figure 7. See APPENDIX D for additional post processing procedures.

RGO and MAB databases contain 20,000 and 13,000 phones respectively. For comparison, Clark et al. (2007) evaluated the Festival text-to-speech synthesis on four databases ranging from 14,000 phones to 175,000 phones. The smallest database tested was collected using a Carstens AG100 (RGO and MAB were collected with the more recent Carstens AG500). The authors noted that the articulatory instrumentation reduced the segmental quality of the synthesized speech and made the voice sound “somewhat unnatural.” The authors ultimately concluded that a database with 36,000 phones was the minimum possible size to achieve reasonable performance (mean opinion score of 3 out of 5) for text-to-speech synthesis. Despite this unfavorable outlook, however, the RGO database is the most extensive single-session collection of EMA data. For comparison, the MOCHA-TIMIT (Wrench, 1999) and X-Ray Microbeam (Westbury, 1994) datasets contain 30% and 50% fewer sentences per speaker.

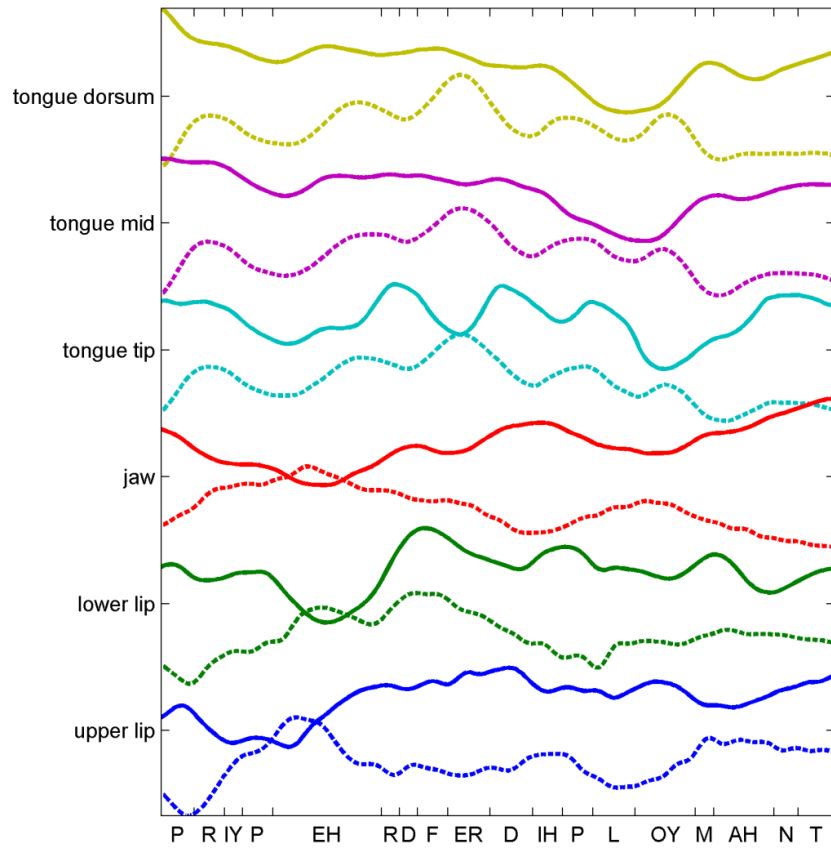


Figure 7 Articulatory trajectories for the phrase “prepared for deployment.” Features are grouped in their x- and y-coordinate pairs; x-coordinates are plotted with a dashed line. The data has been artificially scaled and offset for appearance.

## 2.4 Synthesis and modification of speech

We next review conventional approaches to speech synthesis and modification. The final topic of voice conversion is closely related to accent conversion.

### 2.4.1 Speech synthesis

The ultimate goal for speech synthesizers is to artificially create speech that sounds natural and intelligible, but so far this goal has been elusive. Approaches to speech synthesis can be divided into two broad categories: rule-based (e.g. articulatory and formant) and corpus-based (e.g. concatenative and HMM).

Traditional articulatory synthesizers model the physical act of speaking rather than speech acoustics. Given a model of the vocal tract, the vocal tract transfer function can be calculated. In Figure 8, a cross-sectional view of the vocal tract is approximated using a tube-model; the vocal tract transfer function is then directly calculated using the planar wave equations:

$$\frac{1}{\rho c^2} \frac{\delta(p(x, t)A(x, t))}{\delta t} + \frac{\delta u(x, t)}{\delta x} = \frac{\delta A(x, t)}{\delta t}$$

$$\rho \frac{\delta(u(x, t)/A(x, t))}{\delta t} + \frac{\delta p(x, t)}{\delta x} = 0$$

where  $u(x, t)$  is the air volume velocity and  $A(x, t)$  is the cross sectional area of the tube (Riegelsberger, 1997). Additional refinements can be made by considering source/filter interactions, radiation at the lip, glottal-source characteristics, and acoustic losses of the vocal tract (caused by viscosity, thermal conductivity, and wall vibration). This type of articulatory synthesis produces intelligible, but unnatural sounding speech because the models largely depend on expert guesses (rules).

Another class of articulatory synthesizers attempts to model the articulatory-to-acoustic relationship rather than vocal tract itself. These methods employ a large database of

articulatory/acoustic spectrum pairs to create a mapping from articulatory parameters to the acoustic spectrum. A variety of mapping techniques have been proposed. Shiga and King (2004) proposed a codebook approach that finds a single spectral envelope for each cluster of the articulatory parameters. The estimated spectrum for an unknown articulatory configuration is the envelope associated with its nearest cluster. A related approach (Kaburagi and Honda, 1998) estimates a spectrum using a weighted sum from the nearest codebook entries (weights are inversely proportional to the distance of the corresponding articulatory parameters). Toda et al. propose a statistical articulatory-to-acoustic mapping using Gaussian mixture models (Toda et al., 2008), which was later modified to take advantage of phonetic information (Nakamura et al., 2006). Hidden markov models can also be used for this purpose (Hiroya and Honda, 2004) (Ling et al., 2009).

Formant synthesis is a rule-based approach that directly models speech acoustics rather than the process that creates them. Klatt's formant synthesizer (shown in Figure 9) uses 19 parameters to describe speech, including formant location, bandwidth, amplitude, gain, pitch, and glottal shape (Klatt, 1980). This level of control makes formant synthesizers invaluable for research in speech perception. However, while the resulting speech is highly intelligible, deriving accurate values for the parameters is difficult and time consuming.

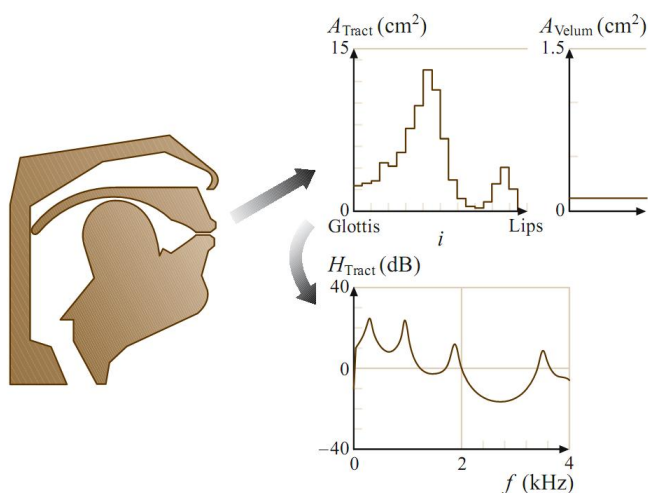


Figure 8 Articular speech synthesis mimics the human speech production process. In this example, a 2-dimensional view of the vocal tract is approximated using a discrete area function  $A_{\text{Tract}}$ . Given this discrete representation, the vocal tract transfer function  $H_{\text{tract}}$  can be calculated from the planar wave equation [reprinted from (Schroeter, 2008)].

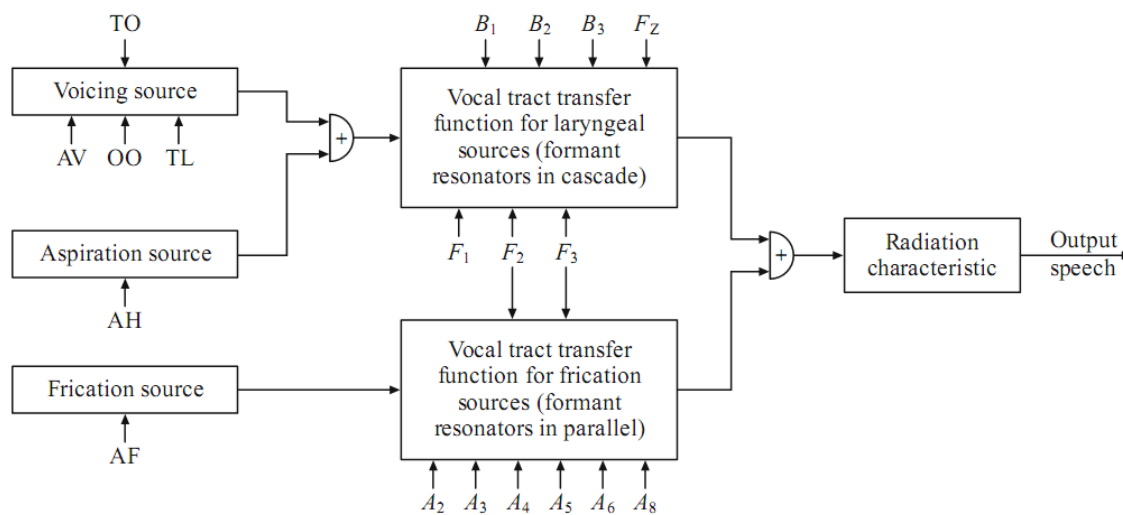


Figure 9 Klatt's hybrid serial/parallel formant synthesizer uses 19 parameters to synthesize speech [reprinted from (Schroeter, 2008)].

Today's state-of-the-art speech systems are based on concatenative or statistical techniques. The concatenative approach works by "gluing" together previously recorded chunks of speech. In theory, the minimum number of sounds a concatenative synthesizer needs is the number of phonemes in the language (about 40 in English). In practice, this does not produce natural speech because it neglects coarticulatory effects. To address the issue of coarticulation, many systems employ diphones (adjacent phoneme pairs) at the expense of increased memory requirements since there are potentially  $40 \times 40 = 1,600$  diphones in English, though 1,000 is a better estimate since not all pairs are used (Syrdal et al., 1995)). For applications that have a limited vocabulary (e.g. telephone directory systems), storage at the word level may be feasible, but this is also not straightforward since words spoken in isolation are quite different from those in conversation (Klatt, 1987). Phone, diphone, and word-based synthesizers that contain only one sample for each unit are known as fixed-inventory synthesizers. On the other hand, unit selection synthesizers store several instances of each unit, thereby improving the chances of finding a well-matched unit. Unit selection needs a large speech corpus to ensure phonetic diversity and prevent the selection of poorly matched units. Details on the mathematical formulation of unit selection will be presented in sub-section 6.3.1. Festival (Clark et al., 2007) is an example of a concatenative synthesizer capable of performing both fixed-inventory and unit-selection based synthesis.

Recently HMM-based synthesizers have gained popularity due to their ability to create fluid speech from a small corpus of speech. This form of synthesis is related to automatic speech recognition (ASR). HMM-based synthesizers describe speech as a sequence of phonemes with each phoneme modeled as a three-state HMM (HMM synthesis leverages many algorithms developed for ASR). Synthesis is performed by combining a sequence of these models and imposing dynamic constraints to estimate spectral features vectors (e.g. MFCC). The result is

convolved with a model of excitation to produce speech. HMM synthesis is perceptually smooth, but of poor voice quality due to the synthetic nature of the excitation signal (Tokuda et al., 2002).

#### **2.4.2 Speech modification**

Speech modification is distinguished from speech synthesis in that it *alters* existing speech rather than creating it. Speech modification typically involves altering the duration and intonation properties of speech. In time scaling, one wishes to change the duration of the speech signal without affecting frequency content (e.g., playing back at a higher speed would reduce the duration of the speech signal but also increase its pitch), whereas in pitch scaling one seeks to change the perceived pitch of the utterance without affecting duration.

The most widely used technique for speech modification (both pitch-scaling and time-scaling) is Pitch-Synchronous Overlap and Add (PSOLA) (Moulines and Charpentier, 1990). PSOLA refers to a family of signal processing techniques that decompose speech into short windows (2-4 pitch periods in length) and later combine them to construct a new (possibly modified) speech signal. Several versions of PSOLA have been proposed in the literature, including Fourier-domain FD-PSOLA, linear-prediction LP-PSOLA, and time-domain TD-PSOLA (Moulines and Charpentier, 1990; Moulines and Laroche, 1995). These algorithms perform comparably under modest pitch-scale and time-scale modification factors, but FD-PSOLA is the most robust to spectral distortion during the pitch modification step. Figure 10 illustrates the basic PSOLA framework which involves (1) decomposing the speech signal into a series of short-time analysis signals, (2) modifying each analysis signal, and (3) combining the modified analysis signals. Further details on this method are provided in the description of our first method of accent conversion (Section 4).



Recently, the parametric encodings used by the STRAIGHT (Kawahara, 1997) and Harmonic plus Noise Model (HNM) (Stylianou, 2001) methods have shown to produce more natural speech than PSOLA. The HNM represents speech as a time-varying harmonic component plus a modulated noise component. The decomposition of a speech signal into these two components provides a robust representation suitable for modification and is capable of producing a highly natural-sounding speech. Similarly, STRAIGHT was developed to provide a flexible analysis-synthesis framework for speech perception research (Kawahara, 1997). STRAIGHT has evolved significantly over the years; the current approach decomposes speech into three parts: 1) fundamental frequency, 2) smoothed spectrogram, and 3) a time-frequency periodicity map that controls the ratio of noise at each frequency (Figure 11). This representation enables easy, high-quality modifications of speech. Pitch modification is performed by updating the fundamental frequency parameter and time modification is performed by duplicating/deleting frames. We use STRAIGHT synthesis in our second method of foreign accent conversion.

### **2.4.3 Voice conversion**

Voice conversion (VC) is the process of modifying speech to alter the perceived identity of a *source* speaker. The desired identity change is usually specified by providing samples of speech from a *target* speaker. VC has applications in speech synthesis and speech perception. As an example, VC can be used to increase the number of available voices in speech synthesis systems. Instead of storing a large amount of data for each voice, a single voice for each gender is stored and then additional voices are stored as transformations of the baseline voice. It has also been used to investigate the acoustic features related to speaker identity and defeat speaker identification systems (Qin et al., 2008). In the entertainment industry, VC may be used to regenerate voices for actors/actresses who are no longer alive.

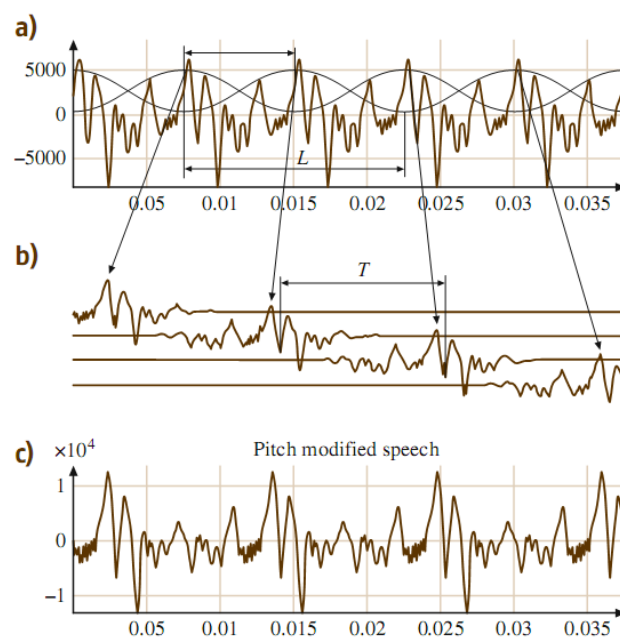


Figure 10 PSOLA modification of speech. The original speech is (a) windowed at pitch synchronous locations, (b) resequenced to simulate a longer pitch period, and (c) added together to give the impression of a lower pitch [reprinted from (Schroeter, 2008)].

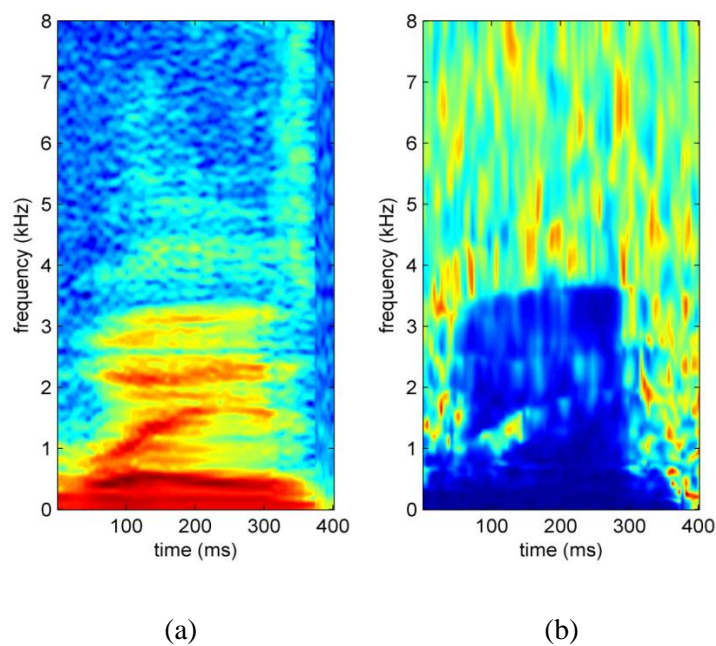


Figure 11 STRAIGHT parameterization of speech consists of (a) smooth spectrogram and (b) aperiodicity signal ( fundamental frequency not shown).

Voice conversion is performed in four basic steps: acoustic modeling, alignment, mapping, and synthesis (Turk and Arslan, 2006). This process is summarized in Figure 12. During the acoustic modeling step, the short-term spectral properties of the speech signal are captured into a low-dimensional feature vector, for both source and target speech signals. Linear Predictive Coding (LPC) is one popular representation as it captures resonances in the vocal tract and can also be used to estimate the glottal excitation signal through inverse filtering<sup>4</sup>. During the alignment step, utterances from the source and the target are time aligned, typically in an automatic fashion by means of Dynamic Time Warping (Abe et al., 1988) or Hidden Markov Models (Arslan, 1999). During the mapping step, a transformation from source to target features is found through machine learning. Common mapping techniques include vector quantization (Abe et al., 1988), neural networks (Narendranath et al., 1995), and Gaussian mixture models (Kain, 2001; Stylianou et al., 1998; Toda et al., 2007). Baudoin and Stylianou (1996) compared these three methods and found Gaussian mixture models to produce the most convincing transformations, but spectral envelopes tends to be over-smoothed. Finally, the estimated target features are synthesized using a method compatible with the acoustic features chosen during the first step (e.g. LPC vocoder, HNM (Stylianou et al., 1998), or STRAIGHT (Toda et al., 2007)). These methods assume that multiple sentences of speech are recorded for both the source and target speakers (typically around 50).

---

<sup>4</sup> LPC coefficients are sensitive to numerical errors and do not have good interpolation properties (Gold and Morgan, 2000; Moulines and Laroche, 1995). For this reason, other acoustic representations have been found to be more suitable for voice conversion, such as Mel Frequency Cepstral Coefficients (Chung-Hsien et al., 2006; Stylianou et al., 1998) or Line Spectrum Frequencies (Hui and Young, 2006; Kain, 2001; Kain and Macon, 1998; Turk and Arslan, 2006).

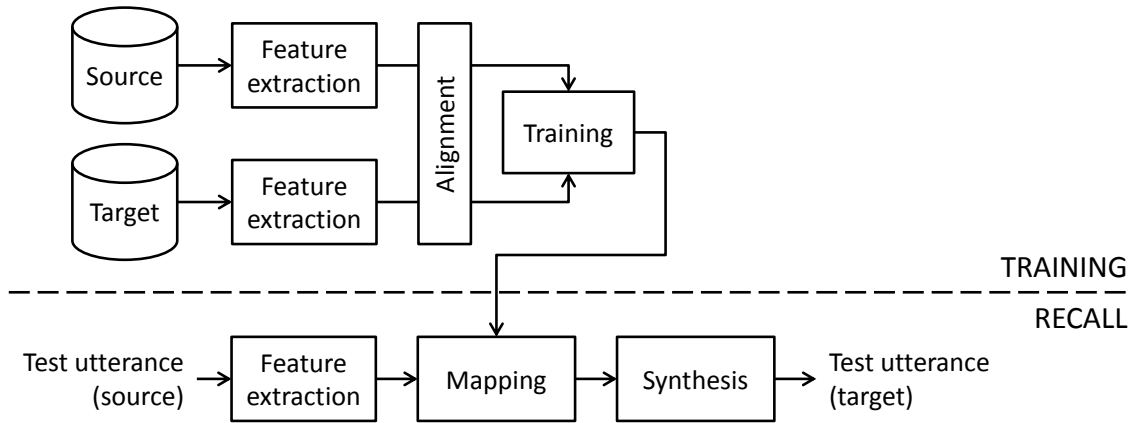


Figure 12 A typical voice conversion system. A mapping is learned from the source to the target speaker in the training phase. Test utterances from the source are transformed in the recall phase.

### 3. LITERATURE REVIEW

This section reviews the current state of accent conversion and previous uses of speech modification to enhance pronunciation training. It also discusses ways to quantify accent, quality, and identity.

#### 3.1 Accent conversion

Accent conversion has grown out of many fields of research including voice and spoken language conversion as well as studies on the perceptual cues of accent. The earliest forms of accent conversion can be found in spoken language conversion (SLC) systems, whose goal is to enable utterances to be created in a language different from that of the source corpus (e.g. create Spanish speech from an English corpus). The primary challenge in SLC is to find phonetic material from the corpus that can be used to realize phones unique to the target language (e.g. Spanish trilled /r/). The simplest solution is to define a mapping to the nearest source phone (e.g. English approximant /r/); such a process can be performed manually using linguistic knowledge or automatically (Teutenberg and Watson, 2006) (Loots and Niesler, 2011).

Campbell (1998) used SLC to synthesize English words from Japanese speech. In this case, five Japanese vowels must be mapped into fifteen English vowels. Allophonic<sup>5</sup> variants of the five Japanese vowels come close to some uniquely English vowels, but the speech database was broadly transcribed without allophonic labels. Campbell proposed to incorporate fine phonetic details representative of the way a native English speaker would produce the utterance. His solution relied on a separate English text-to-speech system to synthesize a native example of

---

<sup>5</sup> An allophone is one of a set of possible sounds used to pronounce a single phoneme. For example, the allophone [p<sup>h</sup>] (as in *pin*) is different from the allophone [p] (as in *spin*), but they are both allophones for the phoneme /p/.

the target utterance. Cepstral values were then extracted from the native utterance and used to select units in the Japanese corpus that had similar cepstral values (Figure 13). MOS evaluation (best=5, worst=1) showed that including cepstral values from a native speaker improved quality (MOS=2.9) over the standard phone mapping approach (MOS=2.3). Though the author did not explicitly measure differences in accent, he noted that utterances created with the proposed method sounded more native than the baseline system. This SLC system provided inspiration for our concatenative foreign accent conversion (Section 6).

Campbell's approach has recently been adapted to perform "accent morphing" (Huckvale and Yanagisawa, 2007). The authors create English-accented Japanese utterances (E) by synthesizing Japanese words with an English TTS. This is accomplished with a special pronunciation dictionary that phonetically spells Japanese words with English phonemes. Accent "morphing" is achieved by altering both prosodic and segmental features of E to more closely match an utterance created with a Japanese TTS (J). Namely, the authors morphed the spectral envelope of E by interpolating line-spectral-pairs and altered E's pitch and rhythm using PSOLA (described in sub-section 4.1). The individual and combined effects of each morph were tested using an intelligibility test, which the authors deemed equivalent to measuring accentedness because English-accented Japanese is less intelligible than native Japanese. Their results show that the segmental and prosodic morphs individually yield a slight improvement in intelligibility. Interestingly, the combined effect was much stronger than that predicted by the individual improvements. Unlike Campbell's SLC, which is limited to the sounds of the source speaker, accent morphing provides a way to create sounds outside of the source corpus. However, if the native TTS voice (J) is dramatically different from the source TTS voice (E), then the morphed voice may not maintain the original identity.

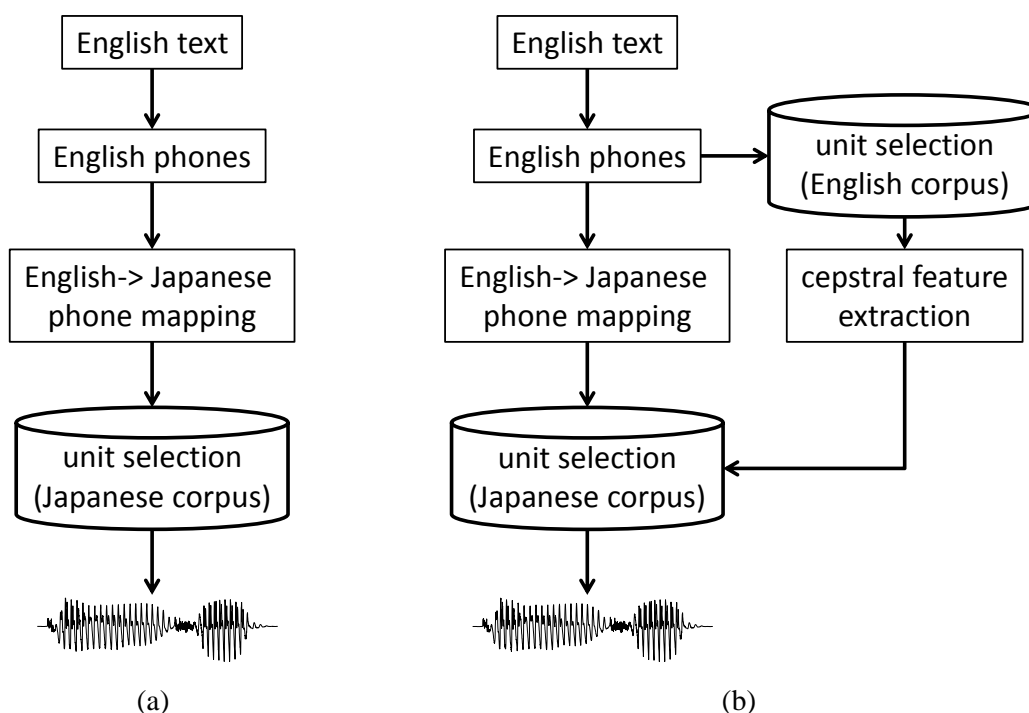


Figure 13 Comparison of two spoken language conversion systems to synthesize English utterances from a corpus of Japanese speech. (a) Basic phone mapping and (b) Campbell's proposed enhancement (Campbell, 1998).

Other approaches modify accent by manipulating the parameters of a formant synthesizer. As an example, Yan et al. (2007) model the formant spaces for vowels in British, Australian, and American accents using two-dimensional (2D) HMMs. A 2-D HMM consists of 1-D HMM in time and a 1-D HMM in frequency. Each state of a 2-D HMM models the distribution of one formant of a phoneme, which can be used to estimate the formant trajectory of a vowel. Accent conversion (of vowels) is performed by altering formant locations, bandwidths, and amplitudes according to the probability distributions obtained from the target accent HMM. Vowel duration is modified with TD-PSOLA according to mean vowel durations from the target accent. Pitch is similarly modified according to broad patterns of pitch trajectories (e.g. initial slope, final slope, pitch range, and overall sentence slope). An ABX test

confirms that accent converted utterances are closer to the target accent than the source accent in about 75% of the cases.

The preceding methods of accent conversion focus on the segmental modification of speech. Yanguas and Quatieri (1999) demonstrated that interchanging the glottal flow derivatives of two speakers could also affect accent. Glottal flow is a measure of volume velocity air flow through the vocal folds (i.e. the “source” in acoustic theory of speech production, subsection 2.3.1). It can be estimated by inverse filtering the speech waveform with an estimate of the vocal tract transfer function. Glottal flow interchange is achieved by convolving one speaker’s glottal flow derivative with another speaker’s vocal tract transfer function over a single pitch period; utterances are created by combining multiple pitch periods with overlap add synthesis. The approach was tested on two sets of speakers in an informal study. The first pair had northern- and southern- American accents, while the second pair had Cuban- and Peruvian-Spanish accents. In both cases, interchanging the glottal flow derivative affected the perceived accent (although no formal results were presented). The authors also noted that modifying the pitch and timing of the speech without changing the glottal flow derivative had a similar but less intense effect.

Most accent conversions systems measure the independent contribution of segmental and prosodic changes to the perception of accent (Felps et al., 2009; Huckvale and Yanagisawa, 2007; Yan et al., 2007; Yanguas et al., 1999). However, the relative importance of segmental and prosodic information is dependent on the L1/L2 pairing. Magen (1998) studied the perceptual contribution of various factors to the perception of Spanish-accented English. She manually edited non-native speech samples to make them sound more native; the investigated factors included syllable structure, vowel quality, consonants, voicing, and stress. Participants rated the original and modified stimuli on a scale from 1 to 7 with 1 corresponding to a native accent. All



factors except voicing were significant, with syllable structure having the largest effect. While it is not practical to manually edit speech to alter its accent, studies like this one can point automatic accent conversion efforts in the right direction.

Two accent conversion systems are studied in this dissertation. Spectral foreign accent conversion (Section 4) is similar in approach to Yanguas and Quatieri (1999). Concatenative foreign accent conversion (Section 6) is most similar to Campbell's spoken language conversion (Campbell, 1998). However, we use real features from a native speaker (rather than estimates from TTS) to find the "most native" diphones from the non-native speaker. Furthermore, we investigate the benefit of using articulatory features to drive the selection.

### **3.2 Speech modification applications in pronunciation training**

During the last two decades a handful of studies have suggested that it would be beneficial for L2 students to be able to listen to their own voices producing native-accented utterances. Nagano and Ozawa (1990) evaluated a prosodic-conversion method for the purpose of teaching English pronunciation to Japanese learners. One group of students was trained to mimic utterances from a reference English speaker, whereas a second group was trained to mimic utterances of their own voices, previously modified to match the prosody of the reference English speaker. Pre- and post-training utterances from both groups of students were evaluated by native English listeners. Their results show students from the second group improved more than those from the first group. More recently, Bissiri et al. (2006) investigated the use of prosodic modification to teach German prosody to Italian speakers. Their results were consistent with those of Nagano and Ozawa (1990), and indicate that the learner's own voice (with corrected prosody) was a more effective form of feedback than prerecorded utterances from a German native speaker. Anecdotal support for the use of accent-conversion is also provided by studies of categorical speech perception and production. In particular, Repp and Williams (1987)

compared the accuracy of speakers imitating isolated vowels in two continua: [u]-[i] and [i]-[æ]. Their results indicate that speakers were more accurate when imitating their own (earlier) productions of those vowels than when imitating vowels produced by a speech synthesizer. Probst et al. (2002) investigated the relationship between the student/teacher voice similarity and pronunciation improvement. Results from this study showed that learners who imitated a well-matched speaker improved their pronunciation more than those who imitated a poor match, suggesting the existence of a user-dependent “golden speaker.” Thus, one can argue that accent conversion would provide learners with the optimal golden speaker: their native-accented selves.

A few computer assisted pronunciation tools have begun to incorporate prosodic-conversion capabilities. These tools allow L2 learners to re-synthesize their own utterances with a native prosody, either through a manual editing procedure or with automated algorithms (Martin, 2004). Proper intonation and stress are especially critical in English because they provide a temporal structure that helps the listener parse continuous speech (Celce-Murcia et al., 1996). Thus, a number of authors have suggested that prosody should be emphasized early on in teaching a second language (Chun, 1998; Eskenazi, 1999). However, speech intelligibility can also degrade as a result of segmental/spectral errors (Rogers and Dalby, 1996), which indicates that both segmental and supra-segmental features should be considered in pronunciation training (Derwing et al., 1998). This suggests that full accent-conversion (i.e., prosodic and segmental) would be beneficial in teaching pronunciation of a foreign language.

### **3.3 Quantifying accent**

A non-native speaker is easily recognized by native speakers because their speech deviates from norms of the native language. These differences are dependent upon the first and second language pairing, e.g. Spanish speakers have difficulty with English vowels and sentence stress (see sub-section 0 for specific examples). Determining the contribution of acoustic

features (e.g. pitch, vowel duration, and formant location/bandwidth) to the perception of accent is an active area of study. In such studies, participants may be asked to evaluate accented speech using an identification task or holistic measure. Identification tasks require participants to classify speech into two or more possible accents. This approach is most useful when the possible dialects are similar (i.e. they share a common first language) (Ikeno and Hansen, 2006; Yan et al., 2007), though it loses sensitivity when differentiating radically different accents (e.g. Spanish accented English verses American English). In turn, holistic measures require participants to rate accented speech according to a Likert scale spanning from their native accent<sup>6</sup> to some non-native accent. This approach has been used to evaluate both prosodic and segmental factors of non-native speech (Magen, 1998; Munro, 1995; Teutenberg and Watson, 2006).

### **3.3.1 Automatic classification of accents**

A variety of methods have also been developed to automatically classify accent, and can be grouped into three categories: methods that model the global acoustic distribution, methods based on accent-specific phone models, and analysis of pronunciation systems (Huckvale, 2007). The first approach models the distribution of acoustic vectors from speakers of a particular accent, e.g. formant frequencies of standard English vowels (Yan et al., 2007). Classification is achieved through pattern recognition, e.g. Gaussian mixture models (GMM) (Chen et al., 2001; Deshpande et al., 2005). Accent-specific phone models have been explored by Arslan and Hansen (Arslan and Hansen, 1996). Their method evaluated words sensitive to accent on separate HMM word recognizers trained for four accents (i.e. English, Turkish, German, or

---

<sup>6</sup> It is also common to use the descriptor “no accent” when referring to an accent that is identical to your own.

Chinese). The accent chosen was the one associated with the HMM that yielded the highest likelihood; their method compared favorably against classification performance by human participants. The final group takes a linguistic approach to accent classification, one which may be more sensitive than methods based on acoustic features (Huckvale, 2007). Barry et al. (1989) compared acoustic realizations *within* a particular speaker. By analyzing systematic differences (or similarities), the authors were able to separate four regional English accents; e.g. Northern English uses the same vowel for “pudding” and “butter,” but American English uses different vowels. Once such phonemic relations are established, it is then sufficient to evaluate the accent of a speaker based on a single sentence that exploits this information. A related approach analyzes a speaker’s phonetic tree to determine accent (Minematsu and Nakagawa, 2000).

### **3.4 Quantifying quality**

This sub-section reviews ways to perceptually or objectively quantify quality. The Comparison Category Rating (CCR) is a perceptual measure of quality that presents participants with pairs of speech samples and asks them to rate the second sample relative to the first using a comparative mean opinion score (CMOS) (3-much better, 2-better, 1-slightly better, 0-about the same, -1-slightly worse, -2-worse, -3-much worse) (ITU-T, 1996). In a related test, Hall (2000) requires participants to listen to groups of 3 stimulus conditions (out of N total conditions) and specify which pair is the most similar and which pair is the most dissimilar. An  $N \times N$  dissimilarity matrix is calculated from these responses (2 points for the most dissimilar pair, 0 points for the most similar pair, and 1 point for the remaining pair). The sum of all responses is analyzed with Multidimensional scaling (Kruskal, 1964) to find a low dimensional embedding where the structure of the data can be visualized. This approach has the potential to give a more complex interpretation of the data than CCR, but it is significantly more demanding on the participants. However, the current ITU recommendation to subjectively evaluate quality is to rate

utterances using a mean opinion score (MOS) (1-bad, 2-poor, 3-fair, 4-good, or 5-excellent) (ITU-T, 1996).

### 3.4.1 Automatic assessment of quality

Objective measures of quality can be broadly described as either *intrusive* or *non-intrusive*. Intrusive measures evaluate the quality of modified speech against the original, high-quality reference speech. The International Telecommunication Union (ITU) recommendation for end-to-end speech quality assessment is P.862, which achieves an average correlation of 0.94 with subjective Mean Opinion Scores (MOS). Such intrusive models are ideal for testing coding or transmission systems because the original, unmodified speech is available for comparison. However, they are not appropriate for voice conversion systems; though a well-defined ground truth exists in this case (i.e. the voice of the native speaker), it is unrealistic to expect a transformed utterance to provide an exact match. For that matter, intrusive models are even more questionable for accent conversion systems because the latter lack a well-defined target.

Non-intrusive measures of speech quality must be used when reference signals are too costly or impossible to obtain, in which case one must predict quality based on the test speech itself. Non-intrusive measures are well suited for testing satellite systems (Jin and Kubichek, 1994), voice over IP, and cell phone networks (Malfait et al., 2006). The most common approach is to create a model of clean speech (e.g. with vector quantization) to serve as a pseudo-reference signal (Jin and Kubichek, 1994). The average distance to the nearest reference centroid provides an indication of speech degradation, which can then be used to estimate subjective quality. Models of the vocal tract (Gray et al., 2000) and the human auditory system (Doh-Suk and Tarraf, 2004) have also been proposed. The prevailing non-intrusive measure is ITU recommendation P.563 (Malfait et al., 2006), which is discussed in detail in sub-section 5.1.

### 3.5 Quantifying identity

Speech is primarily viewed as a source of linguistic information, but secondary types of information (e.g. speaker identity, effort level, emotional state, health, and age) are also encoded as nonlinguistic variations of the basic linguistic message. In an effort to uncover the higher level perceptual processes related to identity, Voiers (1964) asked participants to describe speakers by 49 candidate differential factors (e.g. happy, soft, excited, annoying, or foreign). A factor analysis determined the most significant speaker-discriminating features to be clarity, roughness, magnitude, and animation. As noted by Doddington (1985), there are three problems with this approach: 1) it is difficult for participants to pinpoint the descriptors of a voice that make it unique, 2) these descriptors are not necessarily measurable in terms of acoustic features, and 3) direct use of these factors did not provide effective speaker recognition.

Matsumoto et al. (1973) proposed to infer the psychological acoustic space (PAS) of a subject. Participants were presented with pairs of 24 voice samples and asked to state whether or not they believed the two voice samples to be from the same speaker and indicate their level of confidence on a 3-point scale. A PAS was constructed using multidimensional scaling and the dimensions of the PAS were compared to acoustic features. Matsumoto et al. determined pitch to be the largest contributor to the perception of personal voice quality. Another common approach is to alter various components of speech and evaluate their effect on identity. Analysis-synthesis methods are well suited to this purpose and have been used on multiple occasions (Kuwabara and Takagi, 1991; Lavner et al., 2000). Unfortunately the results of these studies frequently contradict each other. Lavner et al. (2000) note that the search for a single set identifying features may be futile if those features fluctuate among speakers.

The psychoacoustic paradigms mentioned above are designed to discover the perceptual bases for human speaker recognition. Although such information is certainly useful for accent

conversion, they do not provide a quantitative measure of identity (i.e. how does the identity of speaker A relate to the identity of speaker B). For example, voice conversion relies on the ABX test. In this test, participants are asked to listen to three utterances: the first two being the source and target voices (presented in a random order) followed by the transformed utterance. Participants respond with the perceived identity of the final utterance (X), deciding whether it was closer to the first voice (A) or the second (B). Kain (2001) points out that this test is fundamentally flawed since *closer* does not necessarily mean *identical*. Furthermore, it cannot measure cases where the converted utterance sounds like a third speaker. To address this issue, we propose an improved identity test in sub-section 4.3.

### **3.5.1 Automatic classification of speaker identity**

Automatic methods to recognize a speaker from their voice are used in biometric applications (e.g. “voice fingerprint”). Such systems may measure segmental features (e.g. MFCC), pitch, timbre, and jitter (Kuwabara and Takagi, 1991; Matsumoto et al., 1973). Malayath et al. (1997) proposed a multivariate method to separate the two main sources of variability in speech: speaker identity and linguistic content. Namely, the authors used oriented PCA to project an acoustic feature vector (LPC-cepstrum) into a subspace that minimized speaker-dependent information while maximizing linguistic information; this method may also be used for the opposite problem: capturing speaker variability while reducing linguistic content. Lavner et al. (2000) investigated the relative contributions of various acoustic features (glottal waveform shape, formant locations, F0) to the identification of familiar speakers. Their results indicate that shifting the higher formants (F3, F4) has a more significant effect than shifting the lower formants, and that the shape of the glottal waveform is of minor importance provided that F0 is preserved. More interestingly, the study found that the very same acoustic manipulations

had different effects on different speakers, which suggests that the acoustic cues of identity are speaker-dependent.



## 4. SPECTRAL FOREIGN ACCENT CONVERSION

Spectral foreign accent conversion (SpFAC) operates on the principles behind the source-filter model of speech (sub-section 2.3.1): it assumes that the voice quality/identity of speech resides in the source while the linguistic content is contained in the filter. According to this view, the accent of a speaker may be altered by convolving his/her source with the filter of a speaker with a different accent. This is a simplistic view of accent conversion, admittedly with several flaws, but this initial approach to AC served multiple purposes. Namely, it familiarized us with the basics of speech analysis and synthesis (e.g. LPC, pitch tracking, speech modification) and exposed us to the challenges of accent conversion. The remainder of this section provides an overview of the speech modification framework adopted for this work (PSOLA) and describes the segmental and prosodic modifications that make up spectral foreign accent conversion.

### 4.1 Speech modification framework

SpFAC's general framework is based on the analysis/synthesis method of Fourier-domain Pitch-Synchronous Overlap and Add (FD-PSOLA) (Moulines and Charpentier, 1990). The three stages of FD-PSOLA (i.e. analysis, modification, and synthesis) are illustrated in Figure 10.

During the analysis stage, the speech signal  $x(n)$  is decomposed into a series of pitch-synchronous short-time analysis windows  $x(t_a, n)$ . Our implementation (Figure 14) uses a pitch-marking algorithm to estimate instants of glottal closure  $t_a$  (Kounoudes et al., 2002); each analysis window is framed with a Hanning window and transformed into the frequency domain<sup>7</sup>. As a result, all pitch-synchronous short-time spectra  $X(t_a, w)$  are represented with the same length (e.g., 2,048 frequencies in our implementation).

In the modification stage, the short-time spectra and their locations are modified to meet the desired pitch and timing (i.e., those of the American speaker, in our case). First, the short-time spectra are transformed to match the new pitch period, which is equivalent to resampling since we operate in the frequency domain (i.e., spectral compression lowers the pitch and expansion raises it). However, naïve compression of the spectrum also shifts speech formants. For this reason, we first flatten the spectrum with a spectral envelope vocoder (SEEVOC) (Paul, 1981). We also use a spectral folding technique (Makhoul and Berouti, 1979) to regenerate high frequency components that are lost when performing spectral compression (Figure 15 (b)). The resonances of the original spectrum are restored by multiplying the flattened spectrum by the SEEVOC spectral envelope (Figure 15 (c)).

---

<sup>7</sup> Our implementation follows the recommended window length of four times the local pitch period for voiced segments or a constant 10 ms for unvoiced segments (Moulines and Charpentier, 1990).

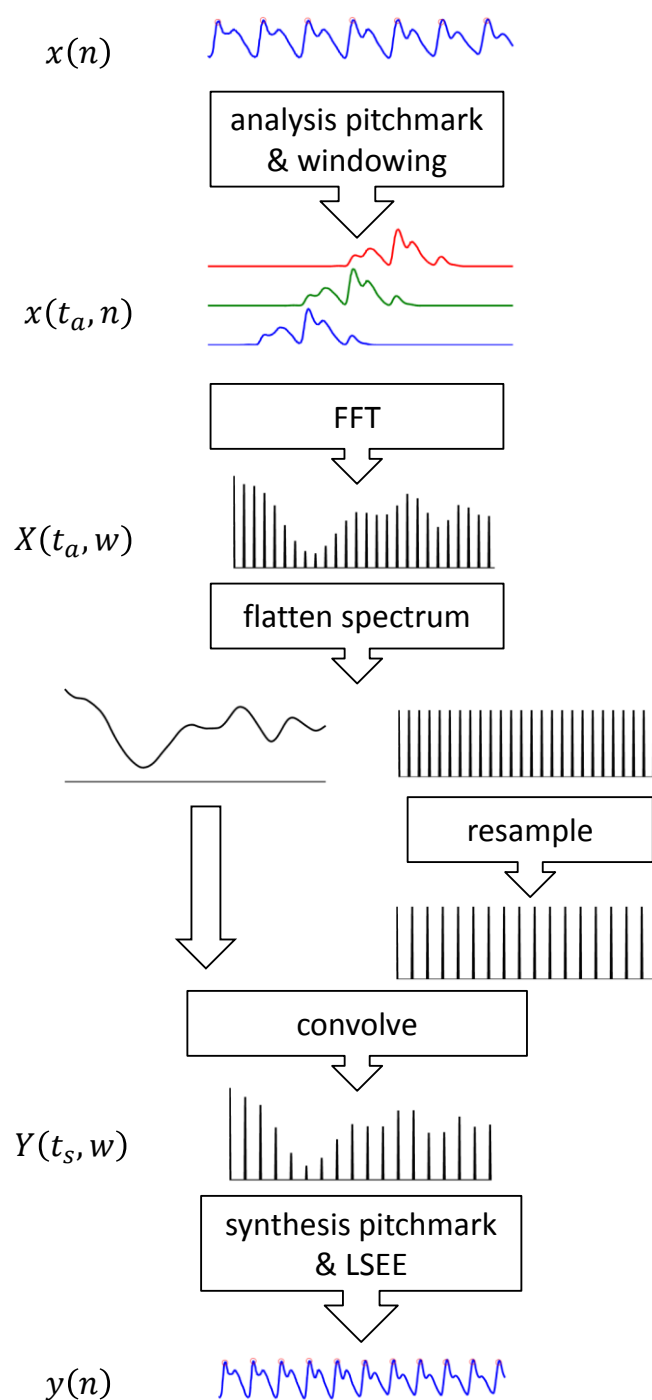


Figure 14 A summary of the FD-PSOLA framework. Speech is divided into pitch-synchronous chunks and transformed into the frequency domain. The spectrum is separated into a spectral envelope and a flattened spectrum. The flattened spectrum is resampled to raise (shown here) or lower the pitch and the spectral envelope is restored. The modified spectrums are then duplicated or deleted as determined by the synthesis pitchmarks to affect duration. Synthesis is performed with the LSEE criterion.

The modified short-time spectra  $Y(t_s, w)$  are copied (i.e., duplicated or deleted) onto the synthesis pitch marks,  $t_s$ , which define the final pitch and timing. For example, duplication of the analysis pitch marks  $t_a$  at the original rate doubles the length of an utterance. On the other hand, duplication of  $t_a$  at twice the original rate doubles the pitch but does not change the length (Figure 16).

The synthesis stage transforms the modified short-time spectra back to the time domain by means of a least-squared-error estimation (LSEE) criterion. Namely, we seek to find the synthetic signal  $y(n)$  whose short-time Fourier transform (STFT) coincides with the target spectra  $Y(t_s, w)$ . However, since  $Y(t_s, w)$  may not be a valid STFT, we seek the *valid* STFT  $\hat{Y}(t_s, w)$  that is closest to  $Y(t_s, w)$  in the least-squares error sense:

$$\sum_{t_s} \int_{-\pi}^{\pi} |Y(t_s, w) - \hat{Y}(t_s, w)|^2 dw$$

This equation can be solved by applying Parseval's theorem; the solution is given in closed form as:

$$y(n) = \frac{\sum_{m=-\infty}^{\infty} w(m-n) F_m^{-1}(n)}{\sum_{m=-\infty}^{\infty} w^2(m-n)} \quad (1)$$

where  $F_m^{-1}(n)$  is the inverse Fourier transform of  $Y(t_s, w)$  at time  $m$  and  $w(m-n)$  is the windowing function (e.g. Hanning) (Griffin and Lim, 1984).

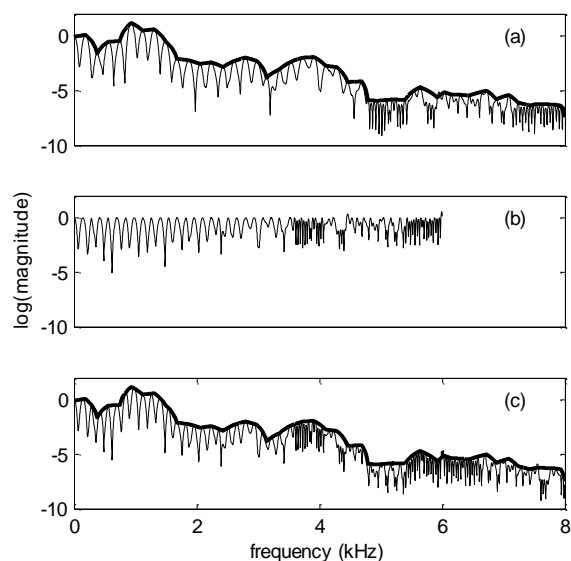


Figure 15 Pitch lowering in the frequency domain. (a) Spectrum of a female vowel /a/ with  $f_0=188$  Hz. (b) The spectrum is flattened and compressed to  $f_0=141$  Hz; notice the spectral hole that occurs at 6-8 kHz. (c) The flattened spectrum in (b) is folded at 6 kHz to fill the hole, and then multiplied by the spectral envelope in (a).

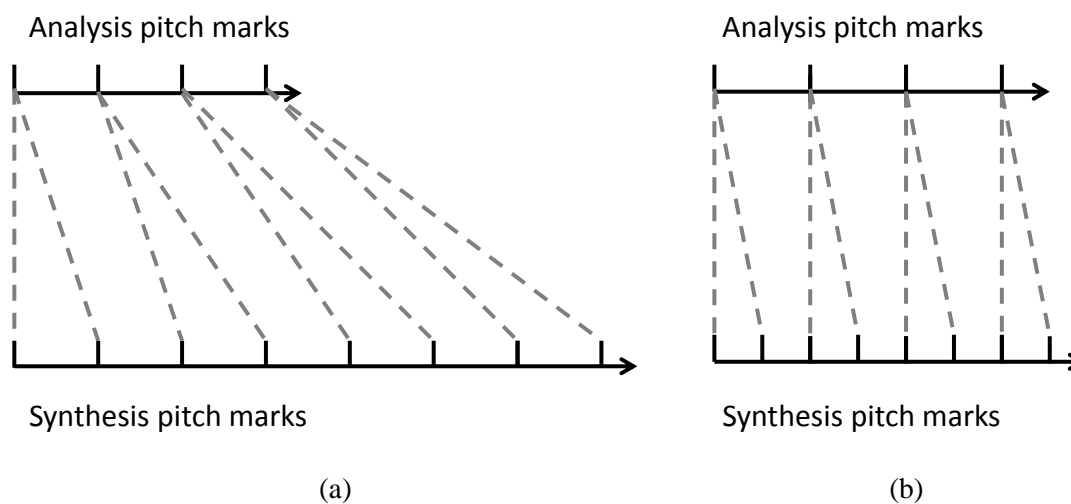


Figure 16 Creating the PSOLA synthesis pitch marks. In (a) the analysis pitch marks are duplicated at the local period rate to double the total duration. In (b) the analysis pitch marks are duplicated at half the local period rate to double the pitch. Similar operations can be used to shorten the duration, lower the pitch, or simultaneously perform any combination of duration and pitch scaling.

## 4.2 Accent conversion

In what follows, we discuss accent conversion in the context of transforming non-native speech to that of a native speaker. We also assume that parallel utterances are available from both speakers. Our accent transformation method proceeds in two distinct steps. First, prosodic conversion is performed by modifying the phoneme durations and pitch contour of the foreign utterance to follow those of the native utterance. Second, the foreign speaker’s spectral envelope (i.e. filter) is replaced by the native speaker’s spectral envelope. These two steps are performed simultaneously in our implementation.

### 4.2.1 Prosodic conversion

To perform *time-scale* conversion, we assume that the native and non-native utterances have been phonetically segmented by hand or with a forced-alignment tool (Young, 1993). From these phonetic segments, the ratio of native-to-foreign durations is used to specify a time-scale modification factor  $\alpha$  for the foreign speaker on a phoneme-by-phoneme basis; as prescribed by Moulines and Laroche (1995), we limit time-scale factors to the range of  $\alpha=[0.25, 4]$ .

Our *pitch-scale* modification combines the pitch dynamics of the native speaker with the pitch baseline of the foreign speaker. This is achieved by replacing the foreign pitch contour with a transformed version of the native pitch contour. For this purpose, we first estimate average pitch values  $(\mu_F, \mu_N)$  for the foreign and native speakers from several hundred utterances. Next, we define a piecewise-linear time-warping,  $\Psi_{FN}(f(t))$ , to align foreign and native utterances at phoneme boundaries. Finally, given pitch contours  $f_0^F(t)$  and  $f_0^N(t)$ , we define a pitch-scale factor  $\beta$  as:

$$\beta(t) = \frac{\Psi_{FN}(f_0^N(t)) + \mu_F - \mu_N}{f_0^F(t)} \quad (2)$$

where we also limit pitch-scale factors to the range of  $\beta=[0.5, 2]$ . This process allows us to preserve speaker identity by maintaining a reasonable pitch baseline and range (Compton, 1963; Sambur, 1975), while acquiring the pitch dynamics of the native speaker (Arslan and Hansen, 1997; Munro, 1995; Vieru-Dimulescu and Mareüil, 2005). Once the time-scale and pitch-scale modification parameters ( $\alpha$ ,  $\beta$ ) are calculated, standard FD-PSOLA is used to perform the prosodic conversion.

#### **4.2.2 Segmental conversion**

The segmental conversion stage assumes that the glottal excitation signal is largely responsible for voice quality, whereas the filter contributes to most of the linguistic information. Thus, our strategy consists of combining the spectral envelope (filter) of the native speaker with the foreigner's glottal excitation. FD-PSOLA allows us to perform this step in a straightforward fashion. As illustrated in Figure 17, we achieved this by multiplying the foreigner's flat spectra by the native's envelope rather than by the foreigner's envelope. In order to reduce speaker-dependent information in the native's spectral envelope, we also perform Vocal Tract Length Normalization (VTLN) using a piecewise linear function defined by the average formant pairs of the two speakers<sup>8</sup> (see Figure 18) (Sundermann et al., 2003). The result is a signal that consists of the foreigner's excitation and the native's spectral envelope normalized to the foreigner's vocal tract length. The result of these transformations is shown in Figure 19.

---

<sup>8</sup> Formant locations were estimated with PRAAT (Boersma and Weenink, 2007) over the entire corpus.

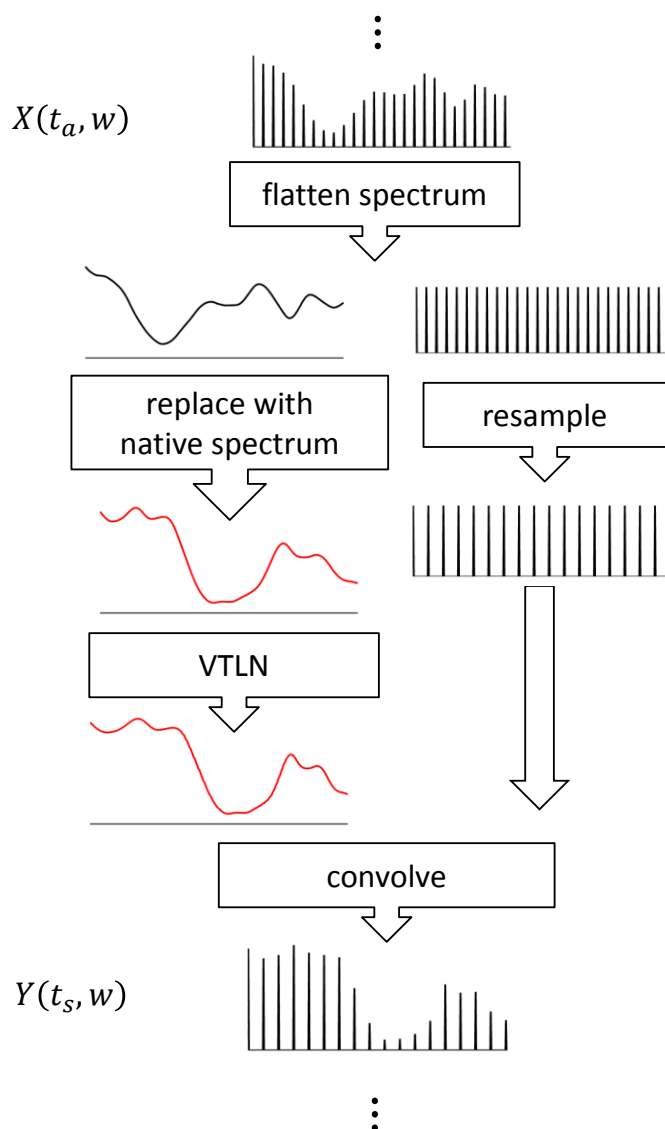


Figure 17 Modifying the FD-PSOLA framework (Figure 14) for the segmental conversion. The segmental conversion replaces the foreign speaker's spectral envelope with the native speaker's spectral envelope after vocal tract length normalization (VTLN). This figure only shows the portion of FD-PSOLA that is modified.



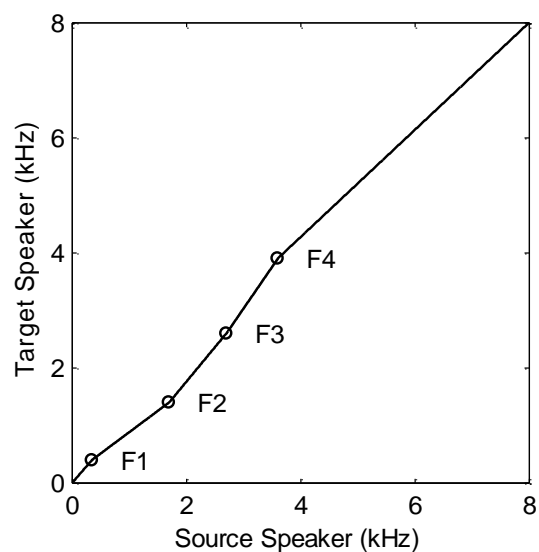


Figure 18 VTLN frequency mapping is created by linearly interpolating between average formant locations for the foreign and native speakers. This physically-motivated transformation preserves acoustic cues associated with the vocal tract length of the foreigner.

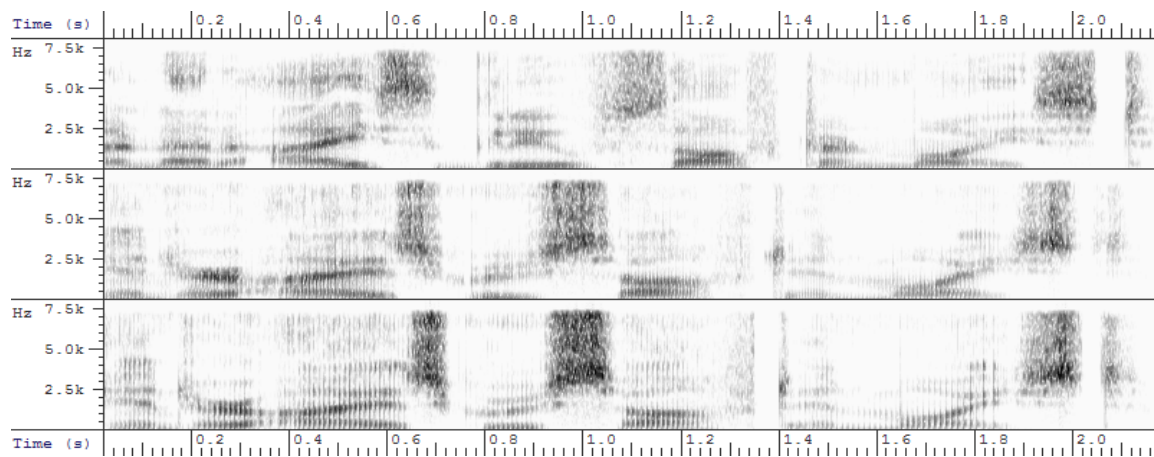


Figure 19 Wideband spectrograms of the utterance “...and her eyes grew soft and moist.” From top to bottom—foreigner, foreigner with prosodic and segmental transformation, and native.

### 4.3 Experimental

To test the proposed method we selected a foreign and a native speaker from the CMU\_ARCTIC database (Kominek and Black, 2003). Given that our participants were native speakers of American English, utterances from *ksp\_indianmale* were treated as the non-native speaker and utterances from *rms\_usmale2* were treated as the native speaker. To establish the relative contribution of segmental and prosodic information, these two factors were manipulated independently, resulting in three accent conversions: prosodic only, segmental only, and both. Original utterances from both foreign and native speakers were tested as well, resulting in the five stimulus conditions shown in Table 4.

Table 4  
Stimulus conditions for perceptual studies.

#	Condition
1	Foreign utterance
2	Foreign w/ prosodic conversion
3	Foreign w/ segmental conversion
4	Foreign w/ prosodic & segmental conversion
5	Native utterance

We were interested in determining (1) the degree of reduction in foreign accent, (2) the extent to which the identity of the original speaker had been preserved, and (3) degradations in acoustic quality. Perceptual evaluation consisted of three independent experiments:

- *Acoustic quality.* Following (Kain and Macon, 1998), participants were asked to rate the acoustic quality of utterances on a standard MOS scale from 1 (bad) to 5 (excellent). Before the test began, participants listened to examples of sounds with various accepted MOS values.

- *Foreign accent.* Following (Munro and Derwing, 1994), participants were asked to rate the degree of foreign accent of utterances using a 7-point Empirically Grounded, Well-Anchored (EGWA) scale (0=not at all accented; 2=slightly accented; 4=quite a bit accented; 6=extremely accented) (Pelham and Blanton, 2007).
- *Speaker identity.* Following (Kreiman and Papcun, 1991), participants listened to a pair of linguistically different utterances, and were asked to (i) determine if the two sentences were produced by the same speaker, and (ii) rate their confidence on a 7-point EGWA scale. These two responses were converted into a 15-point perceptual score (Table 5). To prevent participants from using accent as a cue to identity, utterances were played backwards. This removes most of the linguistic cues (e.g., language, vocabulary, and accent<sup>9</sup>) that may be used to identify a speaker, while retaining the pitch, pitch range, speaking rate, and vocal quality of the speaker, which can be used to identify familiar and unfamiliar voices (Sheffert et al., 2002).

Table 5  
Combined score for identity ratings.

Value	Equivalent meaning
0	Same speaker, very confident
6	Same speaker, not at all confident
7	N/A
8	Different speaker, not at all confident
14	Different speaker, very confident

<sup>9</sup> In (Munro et al., 2003), subjects correctly identified non-native speakers above chance level from reverse speech. However, the authors concluded that this was not due to phonological or prosodic cues, but instead to a long term property of speech (i.e. voice quality). This is consistent with our use of reverse speech to measure identity.

Participants for the subjective tests were recruited from the undergraduate pool maintained by the Department of Psychology at Texas A&M University. All participants were native speakers of American English and had no hearing or language impairments. The audio stimuli were presented via headphones. Thirty-nine students participated in the accent-rating test and forty-three students participated in the quality-rating test. Participants rated 100 utterances (consisting of the same twenty sentences for each of the five conditions) in both of these tests. The identity test was performed with both normal (forty-three participants) and reverse speech (sixty-six participants) to measure the extent to which participants used accent as a cue to identity. Both forms of the identity test required participants to listen to 60 pairs of utterances<sup>10</sup>.

#### 4.3.1 Results

The results are analyzed using a two-way analysis of variance<sup>11</sup> (ANOVA) with the two factors being the prosodic and segmental transformations. Results from the accent ratings are summarized in Figure 20 (a). The original recordings from the foreign speaker received the highest average subjective accent rating (4.85), while native recordings scored the lowest (0.15). The main effect of the segmental transformation was significant,  $F(1,76)=343.03$ ,  $p<0.001$ , indicating that listeners detected a noticeable difference between the original foreign speaker (4.85) and those undergoing the segmental conversion (1.97). Neither the main effect of prosody nor interaction effects were significant.

---

<sup>10</sup> All possible condition pairings can be expressed as a 5×5 matrix. To ensure that all pairs were sampled with the same frequency, diagonal elements in this matrix (i.e. same-same pairings) were sampled twice as often as off-diagonal elements, thus leading to 60 pairs (= (25+5) × 2 repetitions).

<sup>11</sup> A two-factor ANOVA allows us to test the significance of two independent variables (i.e., prosodic and segmental transformations) in the four conditions containing foreign excitation: foreign [0,0], prosodic [1,0], segmental [0,1], and both [1,1]. Results are reported as  $F(df_A, df_{error}) = \_, p < x$ , where the F-score for the independent variable A is dependent upon its degrees of freedom ( $df_A$ ) and the degrees of freedom of the error ( $df_{error}$ ). A complete analysis includes an F-score for each independent variable as well for all possible interactions.

Results from the acoustic quality experiment are summarized in Figure 20 (b). Original recordings from the native speaker received the highest average rating (4.84), followed by those from the foreign speaker (4.0); this difference was statistically significant,  $t(19)=-21.42$ ,  $p<0.001$  (two-tailed). Though recording conditions may have been different for both speakers, it is also possible that listeners penalized the “quality” of non-native speech because it was less intelligible. All transformations lowered quality ratings with respect to the original recordings. Two-way ANOVA found all effects significant: main prosodic,  $F(1,76)=48.48$ ,  $p<0.001$ ; main segmental,  $F(1,76)=119.14$ ,  $p<0.001$ ; and interaction,  $F(1,76)=57.31$ ,  $p<0.001$ .

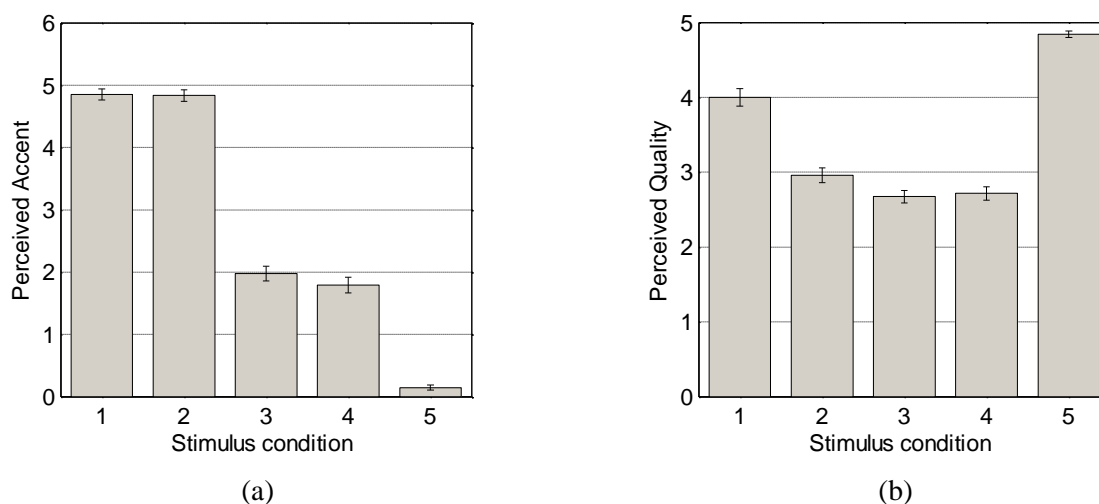


Figure 20 Accent and quality ratings for SpFAC. (a) Accent ratings showing mean  $\pm$  standard error for each stimulus category. The segmental transformation significantly reduced accent. (b) Quality ratings showing mean  $\pm$  standard error for each stimulus category. The transformations significantly reduce quality; note that utterances from the (unmodified) foreign speaker were rated as having lower quality than those from the (unmodified) native speaker. (1=foreign speaker, 2=prosodic transformation, 3=segmental transformation, 4=prosodic and segmental transformations, 5=native speaker).

Results from the identity test yield relative distances (0-14) between the stimuli. Multidimensional scaling (MDS) can be used to find a low-dimensional visualization (e.g., 2D) of the data that preserves pair-wise distances; see (Matsumoto et al., 1973) for a classical use of

MDS in speech perception. Namely, we use ISOMAP (Tenenbaum et al., 2000), an MDS technique that attempts to preserve the geodesic distance<sup>12</sup> between stimuli. Technical details of ISOMAP are included in APPENDIX C. ISOMAP visualizations of the identity tests are shown in Figure 21. In the case of forward speech, samples from conditions 1 and 2 are mapped closely together in the manifold. Thus, this result indicates that the prosodic transformation had only a small effect on the perceived identity of the speakers. On the other hand, non-native utterances (condition 1) and their segmental transformations (conditions 3 and 4) are clearly separated in the ISOMAP manifold. This result suggests that participants heard these utterances as a “third” speaker (note that this type of inference is not possible with the ABX tests commonly used in voice conversion). All samples containing the non-native glottal excitation (conditions 1 through 4) appear to map on a linear subspace that is separate from native utterances (condition 5), which indicates that the former are perceived as being closer to each other than to the native speaker. In fact, by calculating the average Euclidean distance across conditions, we find that this “third” speaker (conditions 3-4) is perceived to be three times closer to condition 1 than to condition 5. Results from the reversed-speech experiment, shown in Figure 21(b), show the distance is reduced between conditions 1/2 and conditions 3/4. This gives weight to the premise that participants in the forward speech experiment identified conditions 3-4 as a “third” speaker primarily due to the association between accent and speaker identity<sup>13</sup>.

---

<sup>12</sup> Thus, ISOMAP assumes that samples exist on an intrinsically low-dimensional surface –a manifold. The geodesic distance is defined as the Euclidean distance between samples measured *over* this manifold. In ISOMAP, the geodesic distance is estimated as the shortest-path in a graph where nodes represent samples and edges indicate neighboring samples.

<sup>13</sup> The spread of the points in the reverse speech condition is likely due to the increased difficulty of the task.

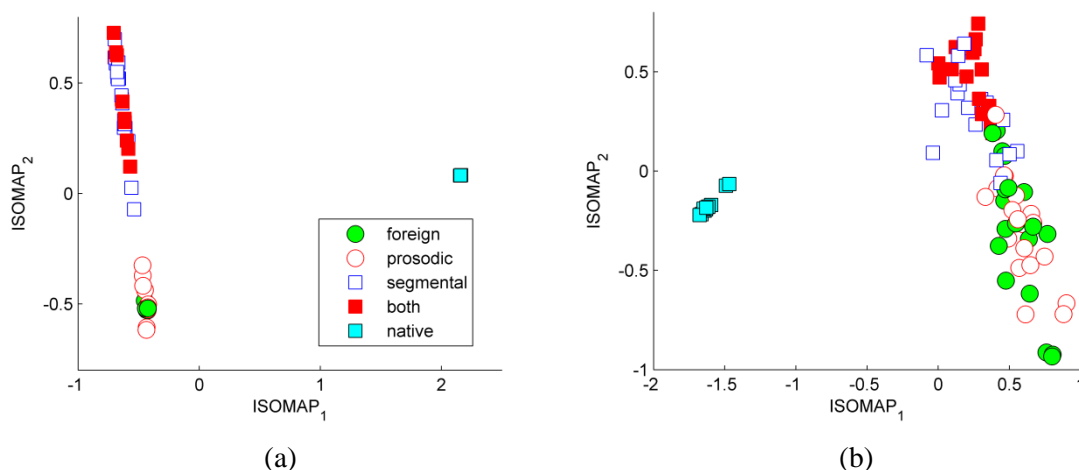


Figure 21 Experimental results from the identity tests. Tight clusters may appear as a single point (e.g., as in the case of native utterances). (a) Forward speech; ISOMAP reveals three distinct clusters: native speaker, non-native speaker (plus prosodic transformation), and a third cluster with the segmental transformations. (b) Reverse speech; ISOMAP reveals only two clusters: native utterances and all other utterances.

### 4.3.2 Discussion

The perceptual results indicate that SpFAC reduces the perceived accent of an utterance. At the same time, this is accomplished at the expense of quality and a change in perceived identity. Although foreign-accented utterances (condition 1) are already perceived as being of lower quality, the technique itself introduces perceivable distortions, as indicated by the lower quality ratings for conditions 2-4. This could be attributed to several factors, including segmentation/alignment errors, voicing differences between speakers, and phase distortions that result from combining glottal excitation with spectral envelope from different speakers. The accent ratings seem to underplay the importance of prosody when compared with other studies (Jilka and Möhler, 1998; Nagano and Ozawa, 1990). One possible explanation for this finding is that the foreigner's prosody sounded relatively native compared to his segmental productions. An alternative explanation is offered by the elicitation procedure in ARCTIC (Kominek and

Black, 2003), since read speech is prosodically flat when compared to spontaneous or conversational speech (Kenny et al., 1998).

Identity tests with forward speech indicate that the segmental transformations (with or without prosodic transformation) are perceived as a third speaker. The distinctiveness of this third speaker is reduced, however, when participants are asked to discriminate reversed speech. One could argue that the emergence of a third speaker on forward speech is merely the result of distortions introduced by the segmental transformation; these distortions are imperceptible when utterances are played backward, which may explain why the third speaker “disappears” with reversed speech. In other words, accentedness and acoustic quality would be confounded in our experiments. This view, however, is inconsistent with the acoustic quality ratings obtained in the second experiment. As shown in Figure 20 (b), quality ratings for condition 2 are similar to those of conditions 3-4, rather than to those of condition 1; if participants had used acoustic quality as a cue in the identification study, condition 2 would have been perceived also as belonging to the third speaker. Thus, our identification experiments with forward and reverse speech suggest that participants used not only organic cues (voice quality) but also linguistic cues (accentedness) to discriminate speakers. This further suggests that something is inevitably lost in the identity of a speaker when accent conversion is performed. After all, would foreign-born public figures (e.g., Arnold Schwarzenegger, Javier Bardem) be recognized as themselves without their distinct accents?

Interestingly, the ISOMAP embedding in both cases (though more clearly with forward speech) can be interpreted in terms of the source-filter theory. As shown in Figure 21, the first dimension separates samples in condition 5, which uses the native glottal excitation, from samples in the remaining conditions, which use the non-native glottal excitation. In contrast, the



second dimension separates samples in conditions 1-2, which employ the non-native filter, from samples in conditions 3-5, which employ the native filter.

#### **4.4 Conclusion**

The preceding method of accent conversion is based on the assumption that accent is contained in the prosody and segmental components of an utterance, whereas speaker identity is captured by vocal tract length and glottal shape characteristics. Our method employs FD-PSOLA to adapt the speaking rate and pitch of the foreigner towards those of the native, and a segmental transformation to replace the spectral envelope of the foreigner with a normalized spectral envelope of the native. This is a somewhat naïve approach to accent conversion because VTLN cannot remove all identity cues associated with the native speaker's spectral envelope. Nevertheless, SpFAC reduced accent by 60% and the results of the reverse speech identity test confirm that the foreign speaker's voice quality was maintained. The next section proposes means to automatically evaluate accent conversion.

## 5. OBJECTIVE MEASURES OF ACCENT CONVERSION

This section presents objective measures of acoustic quality, foreign accent, and speaker identity that are consistent with perceptual evaluations. Such measures would be invaluable in a number of scenarios. As an example, the ability to objectively rate synthesized utterances may be used to search and fine-tune parameters in accent conversion systems – our immediate motivation. Objective measures may also be used in computer assisted pronunciation training (CAPT) to match the voice of the L2 learner with a suitable voice from a pool of native speakers, or to provide feedback to the learner, which is a critical issue in CAPT (Hansen, 2006; Neri et al., 2002).

A major motivation for this work was reducing development time. The ability to evaluate converted utterances in a rapid, unbiased manner is extremely useful for research and development in foreign accent conversion. Time invested in developing these objective measures is quickly returned through time saved by more rapid prototyping and parameter tuning. Admittedly, intermediate development steps are rarely evaluated by formal listening tests, but sidestepping subjective evaluations (even informal ones) is necessary to be able to perform an online optimization of parameters.

### 5.1 Acoustic quality

We adopt ITU recommendation P.563 for our objective measure of quality. It has previously been used for testing satellite systems (Jin and Kubichek, 1994), voice over IP, and cell phone networks (Malfait et al., 2006). The algorithm operates in three stages: preprocessing, distortion estimation, and perceptual mapping. During preprocessing, an additional version of the speech is created by filtering it with a response similar to the properties of a standard telephone.

A third version is filtered with a fourth-order Butterworth high-pass filter with a 100Hz cutoff. Finally, voiced areas are detected using ITU recommendation P.56.

P.563 makes use of several distortion measures to identify the various types of distortions that may be present in the signal. The first measure, based on speech production, approximates the vocal tract area function by transforming the coefficients of an eighth order pitch-synchronous LP analysis into an area function with eight tubes (Gray et al., 2000). These eight tubes are then divided into three groups: front, middle, and rear cavity (corresponding to tubes 1-3, 4-6, and 7-8). Sudden changes in the areas of any of the three cavities indicate the presence of distortion. A second measure of distortion simulates an intrusive quality measure with a reference signal being provided by a speech reconstruction module. The module is designed to remove or modify noise in the distorted speech signal. The reconstructed speech is then compared with the distorted speech using a psychoacoustic model similar to the one found in ITU P.862 (sub-section 3.4.1); this step measures the amount of distortion removed by the speech reconstruction module. The final measures of distortion include estimation of SNR and detection of robotization, temporal clipping, and signal correlated noise.

The final stage, perceptual mapping, takes the above measures of distortion and calculates the final MOS score with a classifier followed by a regression model. The classifier identifies which of seven types of degradations are most likely to be present (i.e. robotization, interruption and clipping, signal correlated noise, low SNR, unnaturally low pitch, unnaturally high pitch, or [default] general distortion). Quality is then estimated on a standard MOS scale using a regression model trained on examples from that class.

P.563 shows an average correlation of 0.85 with subjective MOS (Malfait et al., 2006). The ITU further recommends that, when evaluating a system, multiple speech files be tested and their scores averaged. Despite the fact that P.563 is not intended to measure the quality of accent

transformed speech, we find that it yields reasonable results when at least twenty sentences are averaged per condition.

## 5.2 Foreign accent

The objective measure of accent is related to a method for accent classification based on automatic speech recognition. In Arslan and Hansen (1996), separate HMM word recognizers were trained for four accents (i.e. English, Turkish, German, and Chinese). An utterance was classified to be the accent of the HMM that yielded the highest likelihood. In our method, however, we evaluate a test utterance on a continuous speech HMM (trained on acoustic models from native speakers of American English), and the match score is used as an estimate of its degree of *nativeness*. The primary advantage to this approach is that, as long as the desired accent remains American English, then one need not train a separate HMM for an arbitrary foreign accent.

Our implementation of the continuous speech recognizer is based on the freely-available Hidden Markov Model Toolkit (HTK) (Young, 1993). We call the HTK's general purpose word recognizer "HVite" with flags to specify forced alignment (-a) and to output the calculated log likelihoods (-o); details of the procedure may be found in the HTK book (Young et al., 1995). This step aligns a standard American English pronunciation of the transcript to the provided speech sample. The objective score for an utterance is given as the median value of the phoneme-level match scores (i.e. negative log likelihood) contained in the label file, excluding those associated with silences. As a result, utterances that are more native will have smaller values.

To test the effectiveness of this measure against multiple dialects and accents of English, we selected 108 speakers from the IDEA database (Meier and Muller, 2009). Each of these speakers read the "*Comma gets a cure*" passage and belonged to a country with at least four such

speakers, for a total of thirteen countries<sup>14</sup>. We employed spectral subtraction (Boll, 1979) to reduce background noise levels, which can vary significantly from speaker to speaker in the IDEA corpus. The resulting accent scores, summarized in Figure 22 (a), show a separation between countries that speak English as a first language and those that do not. A two-tailed  $t$ -test<sup>15</sup> found the means of the two groups to be significantly different;  $t(107)=7.16$ ,  $p<0.001$ . Given that the HMM's acoustic models were trained on native speakers of American English, it is surprising that Australia outscored General American (genam), though the difference was not significant  $t(22)=0.16$ ,  $p=0.87$ . In fact, the first country with a significantly different score from Australia was India;  $t(20)=2.22$ ,  $p<0.05$ , the best scoring country where English is not the official language (English is considered a subsidiary official language).

To determine the effect of residual noise on the likelihood scores, we selected seven speakers (with clean recordings) from the CMU ARCTIC dataset (Kominek and Black, 2003). Ten sentences from each speaker were measured with ten levels of additive white noise spanning the range of objective quality measures. Results from one of the speakers (ksp\_indianmale) are shown in Figure 22 (b); a strong correlation<sup>16</sup>  $r(98)=-0.87$ ,  $p<0.001$  between our objective measures of accent and quality indicates that this effect is significant. We adjust all accent scores by the average trend of the seven speakers (120 accent points per quality point) to obtain a measure of accent that was (linearly) independent from acoustic quality.

---

<sup>14</sup> We recognize that there may be multiple dialects within a single country and selected speakers with the same dialect when possible. Namely, we chose Received Pronunciation (RP) from England and General American (genam) from the United States.

<sup>15</sup> The  $t$ -test determines if the means between two groups are significantly different. Results are reported as  $t(df) = \_, p < x$ , where the  $t$ -score is dependent upon the degrees of freedom ( $df$ ); the significance value ( $p$ ) is deemed significant if  $p < 0.05$  (i.e. there is less than a 5% chance that the groups were drawn from the same distribution).

<sup>16</sup> The Pearson product-moment correlation tests the relationship between two variables. Results are reported as  $r(df) = \_, p < x$ , where the magnitude and sign of  $r$  indicate the strength and direction of the relationship;  $r$  ranges from [-1,1].

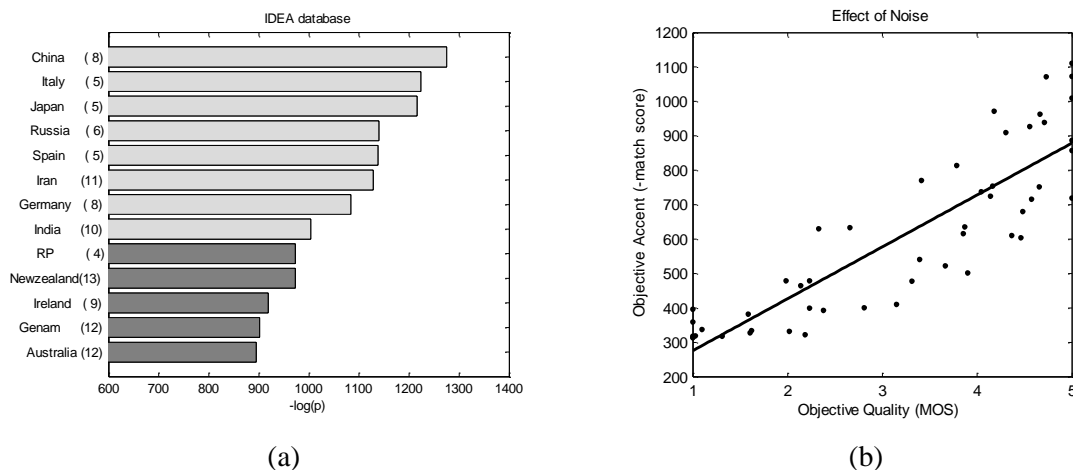


Figure 22 Calibrating the HTK score. (a) Average accent scores for 13 dialects and accents in the IDEA database. Dark colored bars represent countries that speak English as a first language. The number of speakers per country is given in parenthesis to the right of the name. (b) Effect of additive Gaussian noise on the HTK score. Ten sentences with ten levels of noise spanning the full range of MOS values were used for this purpose.

### 5.3 Speaker identity

Our objective measure of speaker identity is based on a signal discrimination criterion (Bishop, 2006). Namely, given a collection of acoustic features for two speakers, we find a projection that maximizes the separability between them by means of Fisher's Linear Discriminant Analysis (LDA). This approach compares favorably against conventional methods for speaker recognition based on GMMs. The primary advantage stems from the fact that LDA is a supervised method, whereas GMMs are unsupervised. GMMs are trained to model the distribution of data in feature space without regard to a feature's discriminatory ability or noise level. LDA, on the other hand, finds the subspace with the highest discriminatory information. This is particularly advantageous when acoustic features are poorly selected or when speakers have broadly overlapping distributions in feature space. In addition, as suggested by Lavner et al. (2000), discriminatory features may change based on the set of speakers; LDA will automatically adapt in such a situation. Moreover, the computational requirements for LDA are also

significantly lower than GMMs; as shown below, a solution is found through a single matrix inversion, whereas GMMs are trained using the fixed-point method of Expectation Maximization (Dempster et al., 1977). Finally, for binary discrimination problems, the LDA solution is a single dimension, which facilitates interpretation. In summary, we find LDA to be a powerful yet efficient solution for determining an objective measure of speaker identity.

We describe LDA for the two-class problem since it will be used to discriminate the foreign and native speakers. The feature vector  $x$  is a vector of acoustic parameters for each speech frame (F0 and 13 MFCCs in this work). LDA seeks a linear projection vector  $w$  such that the projected data  $y = w^T x$  maximizes the distance between classes relative to the variance within each class. It can be shown that, for Gaussian distributed classes with equal covariance, the optimal linear projection is

$$w = S_W^{-1}(m_2 - m_1) \quad (3)$$

where  $S_W$  is the between-class scatter, and  $m_1, m_2$  are the sample mean of the two classes:

$$S_W = \sum_{n \in C_1} (x_n - m_1)(x_n - m_1)^T + \sum_{n \in C_2} (x_n - m_2)(x_n - m_2)^T \quad (4)$$

$$m_1 = \frac{1}{N_1} \sum_{n \in C_1} x_n \quad m_2 = \frac{1}{N_2} \sum_{n \in C_2} x_n$$

In our implementation, one hundred sentences from each speaker are analyzed (in 20 ms frames) to generate a training set. To avoid overfitting, these sentences are different from those later used for testing and do not include any accent conversions. Once the Fisher's LDA solution  $w$  has been computed from training data according to (3), each new test sentence is framed and analyzed to obtain acoustic vector  $x$  and this vector is projected to obtain the objective measure of identity,  $y$ :

$$y = w^T x \quad (5)$$

Each test sentence is then assigned an identity score that corresponds to the average score across its frames. In this way, if a given test sentence sounds more like the native speaker then its identity score will be closer to the scores obtained for the native sentences in the training set.

## 5.4 Experiment

Values for the objective measures described above were calculated on the same 100 utterances used in the subjective evaluation. The success of an objective measure is determined by how well it matches the corresponding subjective measure, so results will be compared to the subjective scores presented in sub-section 4.3.1<sup>17</sup>. The following results were previously reported in Felps and Gutierrez (2010).

### 5.4.1 Results

Figure 23 (a) depicts correlation of the subjective (x-axis) vs. the objective (y-axis) accent measures at the utterance level (i.e. each point represents an utterance from the test set). The low correlation shown in this plot,  $r(98)=0.21$ , is largely caused by differences in linguistic content among the various utterances. This effect can be seen by looking at the subjective scores for the native speaker in Figure 23 (a); these utterances are scored very similarly by human raters, but cover a wide range of objective ratings. We minimize this effect by calculating a single score per condition; calculating the results in this manner improves correlation drastically,  $r(3)=0.997$  (Figure 23 (b)). Both measures also rank speakers in the same order of accentedness, assigning the highest accent to the foreign speaker and the lowest average accent to the native speaker. The objective measure captured the only main effect detected in the subjective scores: the segmental transformation. This indicates that listeners detected a noticeable difference

---

<sup>17</sup> The objective measure of identity yields identical results on normal and reverse speech. Here we compare the objective score to the subjective scores in the reverse condition because this approach was shown to better isolate the relationship between accent and identity (recall the result in Section 4.3.1).



between the accent of the foreign speaker and the segmental conversion, though the effect was more prevalent in the subjective test ( $F(1,76)=343.03$ ,  $p<0.001$ ) than in the objective test ( $F(1,76)=4.48$ ,  $p<0.05$ ).

The utterance level correlations between P.563 and subjective MOS are also low  $r(98)=0.47$  (Figure 24 (a)). The average score per condition yields a higher correlation  $r(3)=0.80$ , which is close to the 0.84 correlation reported in the ITU document (Malfait et al., 2006). Figure 24 (b) clearly shows that the modifications induced by the prosodic and segmental conversions create detectable distortions. A two-way ANOVA found all effects to be significant: main prosodic,  $F(1,76)=11.30$ ,  $p<0.005$ ; main segmental,  $F(1,76)=23.76$ ,  $p<0.001$ ; interaction,  $F(1,76)=5.26$ ,  $p<0.05$ . Some of these distortions can be traced back to instances when the state of voicing differs (e.g. in the word “think,” the native speaker produces the correct /k/ sound, but the non-native speaker produces “thing” with an incorrect /g/ sound).

The objective measure of identity was the only objective measure that yielded a high correlation at the utterance level,  $r(98)=0.94$ , which was further improved using average values per condition,  $r(3)=0.99$  (Figure 25). Such strong results give us confidence that LDA captured the unique, speaker-dependent acoustic features. Both measures discriminated native utterances from the rest, and all accent conversions (segmental, prosodic, both) were rated closer to the foreign speaker than the native speaker. Subjective scores show a main effect for the segmental conversion,  $F(1,76)=103.79$ ,  $p<0.001$ , but no other effects were significant. The objective measure shows a main effect not only for the segmental conversion,  $F(1,76)=114.16$ ,  $p<0.001$ , but also for the prosodic conversion,  $F(1,76)=4.35$ ,  $p<0.05$ .

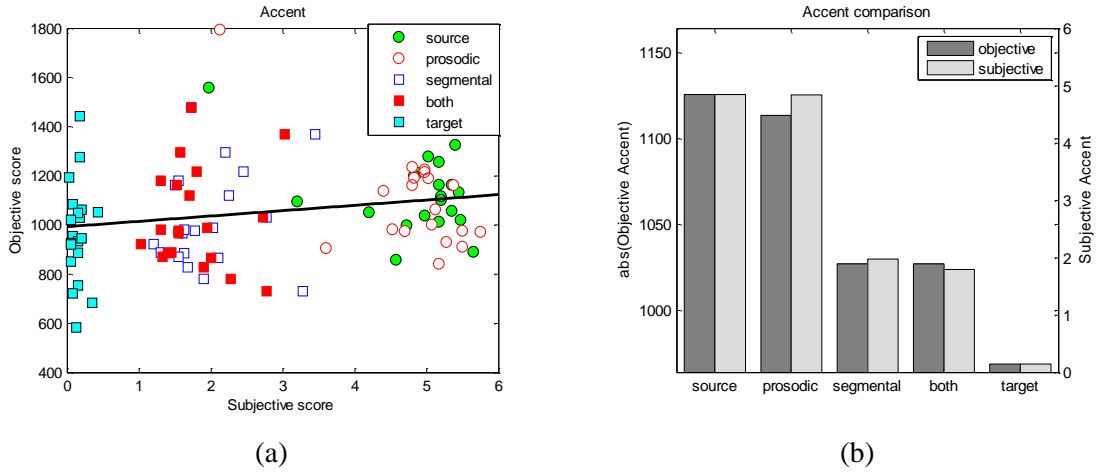


Figure 23 (a) Correlation between objective and subjective measures of accent. (b) Average scores for the two measures across experimental conditions. The objective measure follows the same trend as the subjective scores. Left and right y-axes were aligned to facilitate the comparison.

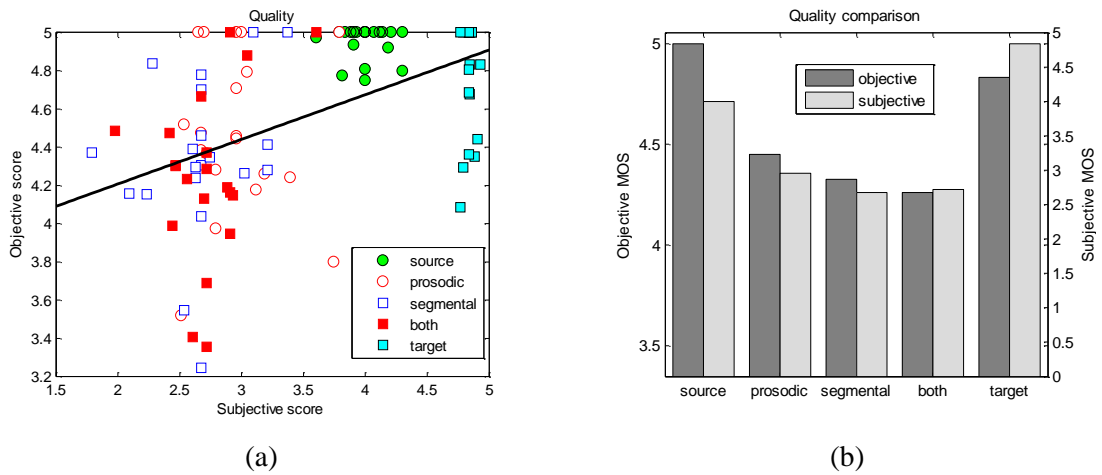


Figure 24 (a) Correlation between objective and subjective measures of acoustic quality. Each sample in the scatterplot represents an utterance. (b) Average scores for the two measures across experimental conditions. The objective measure follows a similar trend as the subjective scores.

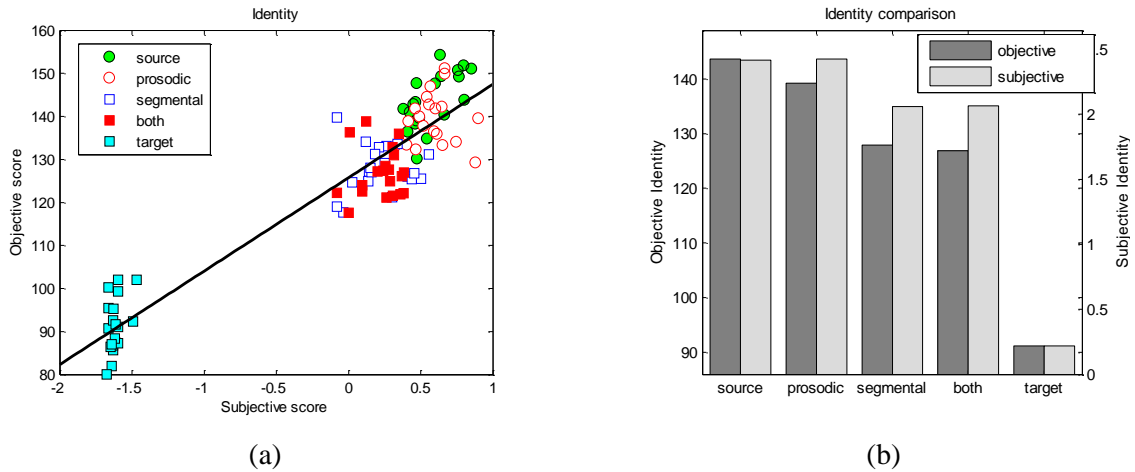


Figure 25 (a) Correlation between objective and subjective measures of identity. (b) Average scores for the two measures across experimental conditions. (b) Experimental results from the identity tests. The objective measure follows the same trend as the subjective scores.

#### 5.4.2 Discussion

We have proposed objective measures that can be used to assess the acoustic quality, degree of foreign accent, and speaker identity of utterances. The three measures show a high degree of correlation across conditions with their corresponding subjective ratings. No attempts were made in our study to match the scales between objective and subjective measures. As an example, the foreign accent ratings Figure 23 (b) have different scales; this issue may be easily addressed by mapping the HTK scores into the 7-point perceptual scale with a regression model or by converting them into an absolute scale by normalizing relative to a corpus of native and foreign-accented speech. However, these extra steps become unnecessary if all one needs are relative measurements, as is the case when optimizing model parameters in an accent conversion system. In this case, it is not the absolute value of the accent measure that is important, but whether it is higher (or lower) than the accent measure for a different set of model parameters; such information is sufficient to guide the optimization engine. Our results also show scaling

differences between objective and subjective measures of acoustic quality, despite the fact that P.563 provides a measure in a MOS scale. These results may indicate a downward bias in our perceptual experiments though participants were provided speech samples with various accepted MOS scores. It seems more likely, however, that P.563 under-penalizes utterances resynthesized with our accent conversion model since P.563 focuses on degradations in narrow-band telephony rather than in speech transformations. Sidestepping these scaling differences, however, our results indicate that the three objective measures are remarkably consistent with perceptual ratings when averaged across sentences.

Unlike the acoustic quality and foreign accent ratings (both objective and subjective), which have a monotonic scale, speaker identity ratings must be interpreted relative to the foreign and native speakers. As an example, consider the identity score for segmental conversions reported by LDA, a value of  $y = 128$  (arbitrary units) averaged across 20 utterances. This value can only be interpreted when compared against the scores for the foreign ( $\mu_F = 143$ ) and native ( $\mu_N = 91$ ) speakers, e.g. segmental conversions are significantly closer to the foreign speaker. When projected on the LDA solution, utterances from our three accent conversions (segmental, prosodic, and segmental + prosodic) lie somewhere between foreign and native utterances, a reasonable result considering that these conversion combine elements from both speakers (glottal excitation, prosody, formants, and vocal tract length). However, it is possible that an utterance may project outside of the bounds defined by the two baseline cases. This suggests that the LDA scores should eventually be mapped into a measure of distance relative to one of the baseline cases. As an example, a radially symmetric kernel of the form  $d = e^{-(y-\mu_F)^2/(\mu_F-\mu_N)^2}$  may be used to transform the LDA projection into a measure that denotes how close an utterance is to the foreign speaker.

Our objective measures have been tailored for accent conversion, but they can be adapted for voice conversion with slight modification of the methods and interpretations. Though the goals of accent conversion and voice conversion with respect to identity are diametrically opposed, LDA is equally useful in both problems. In voice conversion, for instance, a positive result would find the converted utterances projected closer to the target than to the source. Voice conversion does not make a distinction between accent and identity because most methods implicitly model “segmental” accent, and cross-accent conversion is rarely performed (though cross-language conversion is an active research topic (Sundermann et al., 2003)). However, in cases involving source/target pairs with different accents, it is reasonable to assume that a converted voice with the target accent would be preferred over one with the source accent. In such a case it may be worthwhile to measure accent as an additional component of identity. The objective measure of accent may be more relevant in voice conversion if interpreted as a measure of intelligibility. In the case of acoustic quality, P.563 should be as appropriate for voice conversion as it is for accent conversion.

We primarily see accent conversion as a tool to enhance computer assisted pronunciation training (CAPT). In this regards, the objective measures presented in this section may be used to augment such a tool by pairing L2 learners with an appropriate accent donor from a pool of native speakers (see e.g. (Turk and Arslan, 2005) for the “donor selection” problem in voice conversion). Objective measures may also be used to provide feedback to the learner, which is a critical issue in CAPT (Hansen, 2006; Neri et al., 2002). As an example, measures of foreign accent may be used to track the learner’s progress over time and adapt the CAPT tool accordingly, for instance by increasing the complexity of the exercises as the learner improves their pronunciation; these strategies are known as “behavioral shaping” (Watson and Kewley-Port, 1989).

## 5.5 Conclusion

This section proposed objective ways to measure the three most important aspects of accent conversion systems: accent, identity, and quality. Each test was shown to correlate well with perceptual results from the previous section. Subjective tests are more accurate since they rely on human perception, but they are laborious to collect. Certain situations may warrant the use of the proposed objective measures, which automatically estimate scores that correlate well with human perception. The next section describes the second method of foreign accent conversion is based on a concatenative speech synthesizer, which we developed to overcome the limitations of SpFAC (i.e. it alters identity of the foreign speaker and produces low quality speech).

## 6. CONCATENATIVE FOREIGN ACCENT CONVERSION

SpFAC was partially successful for accent conversion; it significantly reduced accent, but it also compromised the quality and identity of the resulting utterances. In this section we propose a concatenative synthesis approach to foreign accent conversion (ConFAC) to overcome these limitations. In concatenative synthesis, short units of speech (e.g. phones, diphones, or words) are combined to create novel utterances. We hypothesize that this approach will offer a higher level of quality than SpFAC and preserve the identity of the non-native speaker because each unit is taken from a corpus of his/her own speech. In a pronunciation training scenario, ConFAC utterances provide realistically attainable targets for second language speakers. Accent conversion is performed by selecting diphones from a database of non-native speech that match the articulatory/acoustic patterns of a native speaker. Our hypothesis is that articulatory features provide a representation of speech that is more speaker independent than acoustic features.

This view is supported by the front cavity hypothesis, which assigns the portion of the vocal tract captured by EMA (front cavity) to be responsible for linguistic information. Anecdotal evidence can be seen when comparing low dimensional vowel spaces for the two speakers used in our study (Figure 26). The acoustic space computed using the first two principal components of Mel-cepstral coefficients for all vowels found in the RGO and MAB corpora. Likewise, the low dimensional articulatory space computed in a similar manner from Maeda parameters. The average phone distance is twice as large in the low dimensional acoustic space as it is in the low dimensional articulatory space for our two speakers. ConFAC is described in three distinct stages: analysis, accent conversion, and synthesis (Figure 27). The analysis stage encodes a non-native utterance into a standard format that is appropriate for concatenative synthesis (i.e. synthesis features). The synthesis features are then modified during the accent conversion stage to match the corresponding features of a native speaker. Finally, diphones are selected from the non-native corpus based on the modified synthesis features. The selected diphones are processed with optimal coupling and spectral smoothing to create an acoustic waveform.

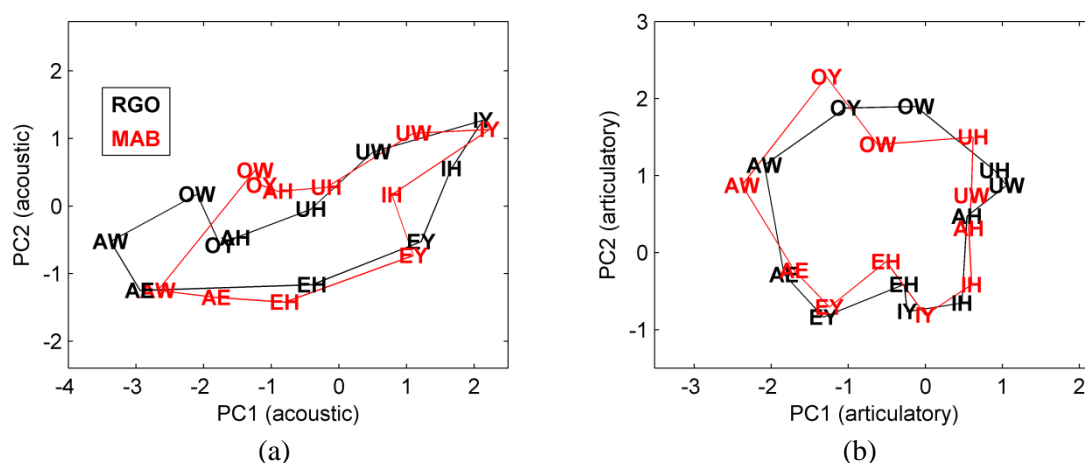


Figure 26 Average location of 11 vowels in a low dimensional (a) acoustic and (b) articulatory vowel spaces. The average distance between corresponding phones for RGO and MAB is twice as close in articulatory space.



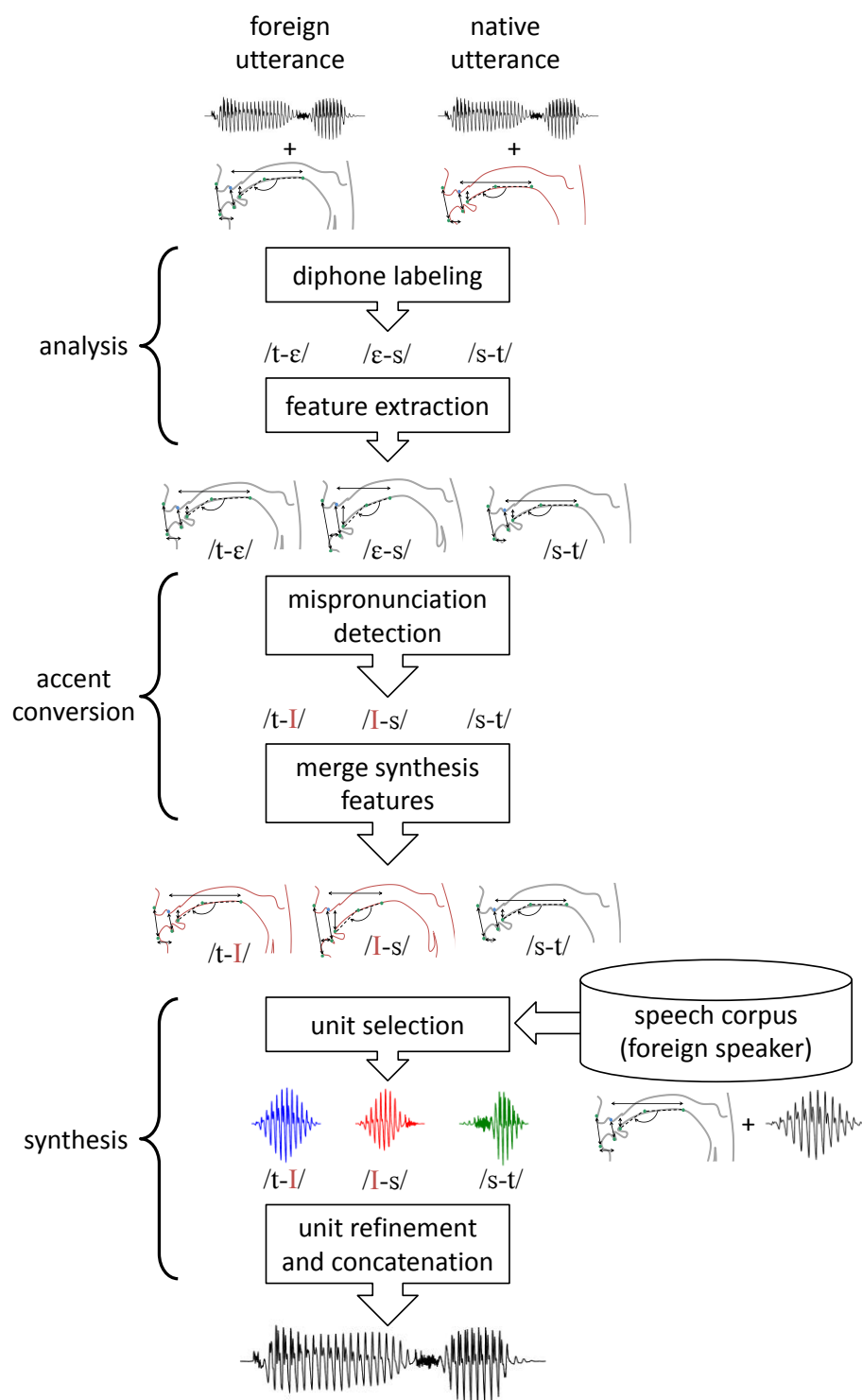


Figure 27 ConFAC system overview for articulatory-driven accent conversion. ConFAC selects diphones from a foreign speaker that match the articulatory patterns of a native speaker. The process to perform acoustic-driven accent conversion is similar except the Maeda parameters are replaced with MFCCs.

## 6.1 ConFAC analysis

The analysis stage is responsible for labeling and extracting features for each diphone in a non-native utterance (Figure 28). The first step is to determine the location and labels of phones in the utterance. This is performed by manually correcting estimates obtained using HTK forced alignment (see sub-section 5.1 for more information). Diphones are then defined from the center of one phone to the center of the following phone. Diphone synthesis is more natural than phone synthesis because diphone boundaries are more acoustically stable than phone boundaries. Next, information about each diphone is encoded into a common format: Arpabet label (e.g. /a-t/), duration (e.g. 100 ms), and temporal features (e.g. pitch, loudness, MFCCs, articulatory trajectories). In order to have a common representation with diphones of different lengths, each temporal feature is sampled at three relative locations: the beginning, middle, and end of a diphone. These features, called *synthesis features*, are compatible with the unit selection synthesizer.

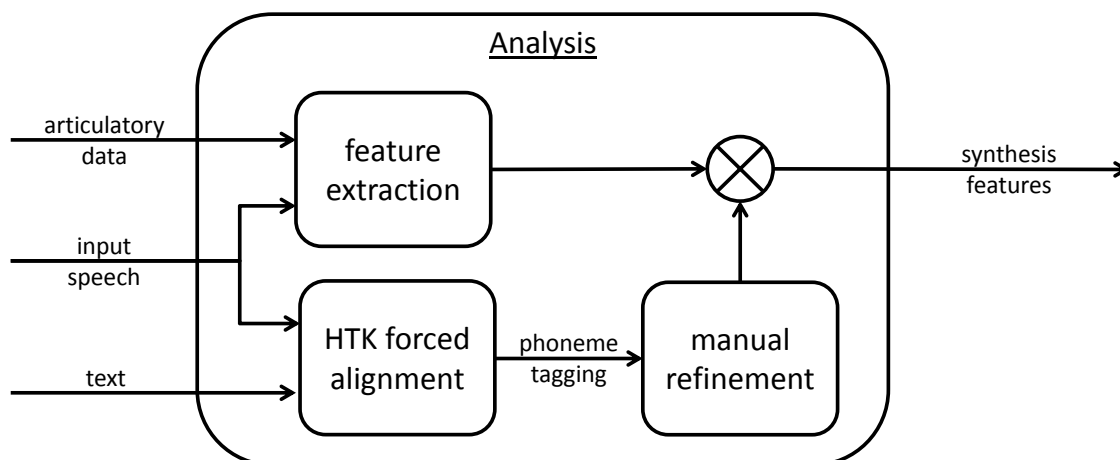


Figure 28 ConFAC's analysis stage encodes an utterance into a format that permits unit selection based synthesis.

## 6.2 Accent conversion

The accent conversion stage modifies the non-native synthesis features to encourage “more native” sounding units to be selected during the synthesis stage. The first step is to identify differences in the native and non-native phonetic sequences. As shown in Figure 29, we calculate synthesis features for native and non-native examples of the same sentence. Differences in the native and non-native phone sequences are specified manually by creating a mispronunciation file to fit the native phone sequence to the non-native utterance. This process is illustrated in Figure 30 for the word “anything.” In this example, the non-native speaker says “any-ting” with a hard /g/. The mispronunciation file registers the native speaker’s phonetic pronunciation to the best fit of the non-native speaker’s actual production. Substitutions, deletions, and insertions are easily detected by comparing the mispronunciation file to the original non-native transcription. Phoneme-level mispronunciations are subsequently registered at the diphone level (a single phone-level mispronunciation affects two diphones). Because sounds in continuous speech are not produced in isolation, ConFAC includes a parameter to define how far a “mispronunciation” can spread to neighboring diphones. The default value is 2 diphones to either side of a mispronounced phone.

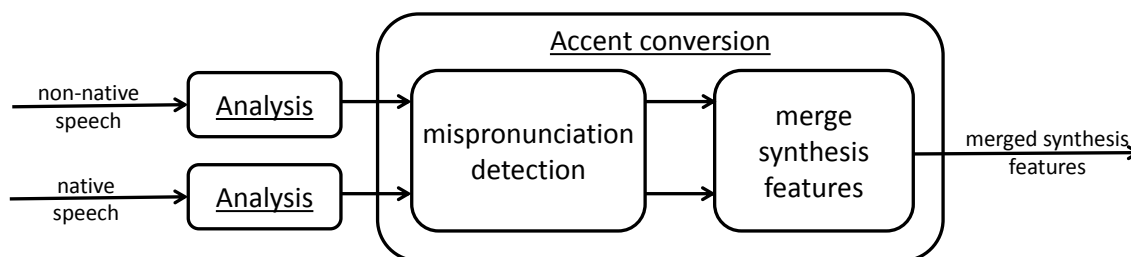


Figure 29 ConFAC’s accent conversion module compares the native and non-native synthesis features. It merges their synthesis features based on the detected mispronunciations.

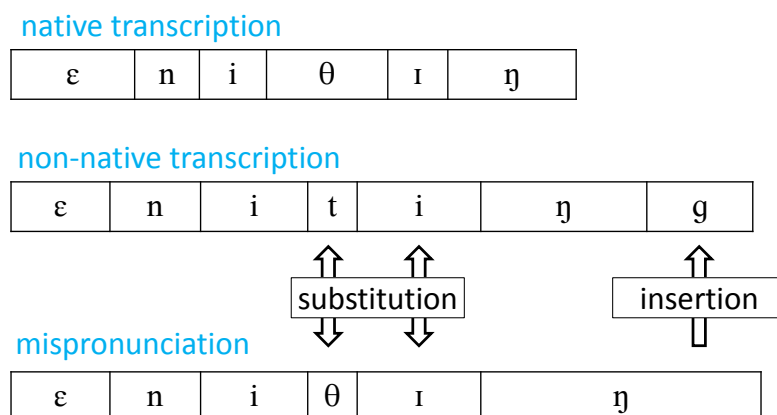


Figure 30 Creating the mispronunciation label file for the word “anything” to identify differences between the non-native and native phonetic sequences. The mispronunciation file specifies the native speaker’s phonetic sequence at the timing of the non-native speaker. In this example, the non-native speaker made two substitutions and one insertion.

Synthesis features from the native and non-native speaker are merged as follows. If a diphone was mispronounced, then non-native synthesis features are replaced with the corresponding native synthesis features. On the other hand, if a diphone was correctly pronounced, the non-native synthesis features associated with suprasegmental properties of speech (e.g. pitch, loudness) are replaced with the corresponding native synthesis features and the remaining features (e.g. MFCC, Maeda) are unaffected. We predict that this process will encourage more “native sounding” diphones to be selected during synthesis. The approach is valid insofar as the native features are within the natural range of non-native features; the next sub-section provides further details on the measures taken to reduce speaker dependent influences.

### 6.2.1 Feature normalization

Speaker normalization is a critical step when performing speech recognition because it increases the similarity of the current speaker to the speakers used during training. A similar step is needed in ConFAC because synthesis is performed on the non-native speaker's corpus using (some) features extracted from a native speaker. Thus, the success of accent conversion for the two feature sets (i.e. acoustic or articulatory) is partially dependent upon each set's ability to provide a speaker-independent representation of speech.

The acoustic features used in this work are Mel Frequency Cepstral Coefficients (MFCC), which are the gold standard in automatic speech recognition and have also been used for speech synthesis (Tokuda et al., 2002). ConFAC calculates 13 MFCCs directly from the STRAIGHT spectrum by first warping the spectrum according to the Mel-frequency scale and then computing the discrete cosine transform. Cepstral mean subtraction is also performed by subtracting the mean and dividing by the standard deviation of each cepstral coefficient. This process reduces the effects of different recording environments and also accounts for differences in long-term voice properties (e.g. spectral slope).

We perform speaker normalization in the articulatory domain by mapping EMA positions to Maeda parameters. Maeda parameters are relative measurements of the vocal tract that explain the majority of articulatory variance (Maeda, 1990). We adopt the EMA to Maeda mapping proposed by Al Bawab et al. (2008) in an investigation of supplemental features for speech recognition (Figure 31). Furthermore, we subtract the mean and divide by the standard deviation of each parameter. The parameters are calculated as follows:

1. Jaw opening distance: the absolute distance from the lower incisor to the upper incisor (origin).
2. Tongue dorsum position: the horizontal displacement between the tongue dorsum and the upper incisor.
3. Tongue shape: the angle created between the three points on the tongue.
4. Tongue tip height: the vertical displacement between the tongue tip and the upper incisor.
5. Lip opening distance: the absolute distance between the upper and lower lips.
6. Lip protrusion: the absolute distance between the midpoint of #1 and the midpoint of #5.

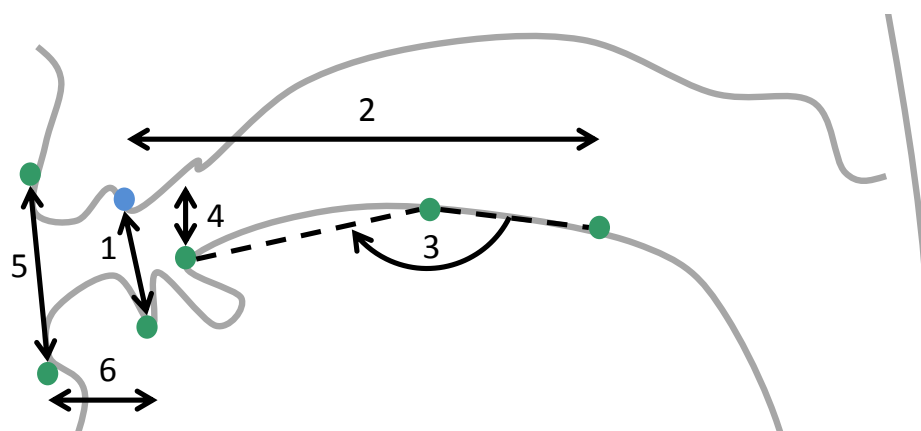


Figure 31 Calculating the Maeda parameters from EMA (x,y) positions. The origin is located on the upper incisor designated by the blue dot. The sixth Maeda parameter is actually measured from the midpoint of #1 to the midpoint of #5. It is vertically offset for clarity.

### 6.3 ConFAC synthesis

ConFAC synthesis creates an acoustic waveform from the synthesis features. It searches the non-native corpus to find diphones that match the synthesis features *and* combine smoothly with each other. The selected diphones are refined to minimize spectral jumps at the diphone boundaries before the synthesizing the final speech waveform. An overview of ConFAC synthesis is presented in Figure 32; details for each step are provided in the next four subsections.

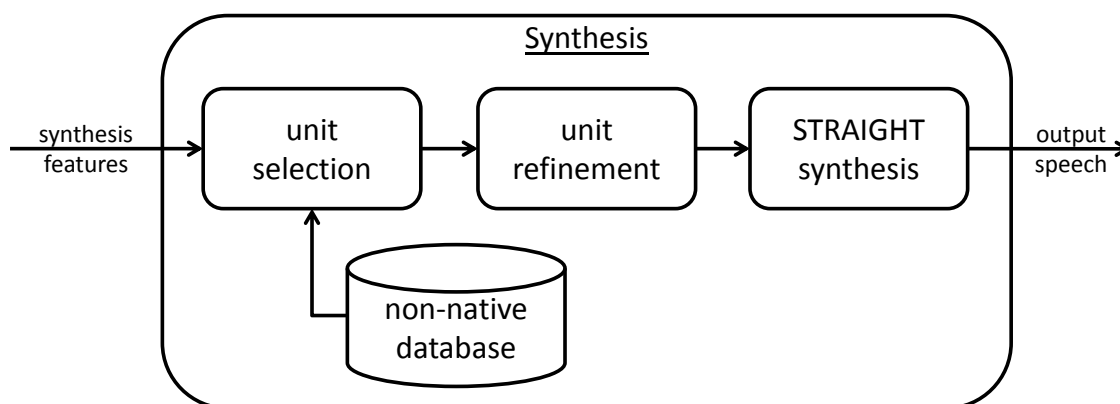


Figure 32 ConFAC's synthesis stage creates an acoustic waveform from the synthesis features.

### 6.3.1 Concatenative unit selection

The primary challenge in concatenative speech synthesis lies in selecting a sequence of units (i.e. diphones) from a large database to yield a natural and intelligible utterance. ConFAC has an additional goal: the utterance should also sound like that of a native speaker. We assume that the synthesis features contain information regarding these three criteria, but the synthesis database will not contain units to perfectly match all the synthesis features. Our task is to find a sequence of units that match the synthesis features as closely as possible and join smoothly together to create natural sounding speech. Hunt and Black (1996) propose a mathematical framework for this problem called unit selection.

In unit selection, a synthesis database is viewed as a state transition network; states represent units and transitions estimate the quality of concatenation between two units (Figure 33). Finding a sequence of units that match the synthesis features is equivalent to finding a path through the network with minimum total cost. Total cost is a sum of the target cost, which penalizes the distance between a potential unit from the database and the synthesis features, and concatenation cost, which acts as a smoothness constraint. The target cost function  $C^t(t_i, u_i)$  is defined as

$$C^t(t_i, u_i) = \sum_{j=1}^p w_j^t C_j^t(t_i, u_i)$$

where  $C_j^t(t_i, u_i)$  is the difference between the synthesis features of the target unit  $t_i$  and a unit from the database  $u_i$ . The importance of the  $j^{\text{th}}$  feature is determined by the weight  $w_j^t$ . Weights are assigned using a standard regression training algorithm (Hunt and Black, 1996). We have previously demonstrated this approach assigns articulatory feature weights similar to those that would be linguistically predicted for certain phones (e.g. a large weight for lip movement in the phoneme /p/) (Felps et al., 2010). The concatenation cost measures the quality of a join between consecutive units  $u_{i-1}$  and  $u_i$ :

$$C^c(u_{i-1}, u_i) = \sum_{j=1}^q w_j^c C_j^c(u_{i-1}, u_i)$$

The quality of a join can be estimated by the distance between MFCCs, pitch, and power at the boundaries of two units. In the context of a state transition network, the target cost serves as the state occupancy cost and the concatenation cost serves as the state transition cost. The total cost for a given sequence (i.e. a path through the network) is the sum of the target and concatenation cost functions for each unit in the sequence:

$$C(t_1^n, u_1^n) = \alpha \times \sum_{i=1}^n C^t(t_i, u_i) + (1 - \alpha) \times \sum_{i=2}^n C^c(u_{i-1}, u_i) + C^c(S, u_1) + C^c(u_n, S) \quad (6)$$

The last two terms in the equation above represent the costs associated with transitioning to and from silence ( $S$ ) at the beginning and end of the utterance;  $\alpha \in [0,1]$  is a user-defined parameter that represents a preference for smooth joins versus accurate target matches. The Viterbi algorithm (Forney, 1973) is guaranteed to find the path through the network that minimizes the total cost. This path represents the sequence of units to be concatenated.



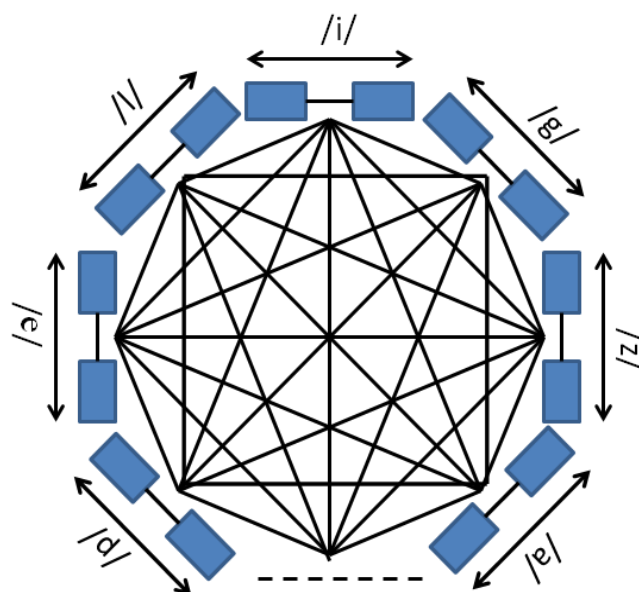


Figure 33 Representing a speech corpus as a state transition network. States (blue boxes) represent individual phones in the database. The state occupancy cost is the target cost and the state transition cost is the concatenation cost.

### 6.3.1.1 Unit selection with a small speech corpus

The recommended corpus size for unit selection is 3-4 hours of phonetically balanced speech (Clark et al., 2007). Performing synthesis without adequate prosodic and phonetic coverage yields speech with spectral distortions and low intelligibility. In comparison, RGO's speech corpus contains approximately 45 minutes of speech (we are not aware of any longer articulatory datasets). Our initial results yielded speech with such low intelligibility that it was rated with a higher accent than the original RGO utterances. We compensate for the reduced size of our speech corpus by modifying unit selection in two ways.

The first modification allows the original RGO diphones to be considered as candidates for synthesis. If these diphones are selected, then the synthesized utterance will be identical to the original foreign utterance. In fact, we can force this to occur by setting the unit selection parameter  $\alpha = 0$ , which gives full weight to the concatenation cost and no weight to the target

cost. Since neighboring units (by definition) have a zero concatenation cost, the Viterbi algorithm will select the original sequence of diphones for a total cost of 0. The resulting synthesis will be distortion-free, but also identical to the original utterance (and therefore have the same accent). As  $\alpha$  increases, unit selection gives greater weight to target costs and the relative benefit of being natural neighbors shrinks (i.e. concatenation cost = 0). The optimal path will now deviate from the original units to include different units from the database. Selecting different units increases the chance of altering accent, but it also increases the chance of introducing distortions. By adjusting  $\alpha$  accordingly, we can effectively control the number of diphones replaced in a given utterance. This number should be as large as possible without introducing distortions. The second proposed modification gives us better control over this factor.

Previously the total cost equation (6) controlled the relative importance given to concatenation costs and target costs using the parameter  $\alpha$ . The second modification replaces  $\alpha$  with a function  $\lambda(\cdot)$  that dynamically controls the percentage of new units to be selected. Let the function  $newUnits(\alpha)$  calculate the percentage of new units selected for a given  $\alpha$ , and the user-defined variable  $perNewUnits$  represent the percentage of new units to be replaced in an utterance. Equation (6) is updated as follows

$$C(t_1^n, u_1^n) = \lambda(perNewUnits) \times \sum_{i=1}^n C^t(t_i, u_i) + (1 - \lambda(perNewUnits)) \times \sum_{i=2}^n C^c(u_{i-1}, u_i) \\ + C^c(S, u_1) + C^c(u_n, S)$$

where

$$\lambda(perNewUnits) = \underset{\alpha}{\operatorname{argmin}} |newUnits(\alpha) - perNewUnits| \quad \alpha \in [0,1]$$

This formulation provides an intuitive way to balance<sup>18</sup> desired the amount of change and the overall level of naturalness without adding significant computation cost; namely, it allows the experimenter to select the percentage of diphones that should be changed.

### 6.3.2 Optimal coupling

Unit selection yields a sequence of diphones to be concatenated, but direct concatenation can lead to harsh distortions caused by discontinuities in the acoustic spectrum. We reduce spectral mismatch between consecutive diphones with two spectral smoothing methods. This sub-section describes optimal coupling, which looks for an acoustically stable place to join two units and the next sub-section describes a spectral interpolation algorithm offering further refinement.

Optimal coupling (Conkie and Isard, 1997) improves the join between two diphones by adjusting the locations of the common boundaries to minimize a measure of spectral distortion (i.e. the right boundary of the left diphone and the left boundary of the right diphone, Figure 34). The cost of a particular join is calculated as follows: let  $l_j^i$  represent a row vector containing the  $N$  ( $N=10$  in our implementation) values of the  $i^{th}$  cepstral coefficient *prior* to the  $j^{th}$  cut-point for the left diphone and similarly  $r_k^i$  for the  $N$  values *following* the  $k^{th}$  cut-point of the right diphone. A line of best fit for the combined vector  $[l_j^i \ r_k^i]$  is specified by the coefficients  $b_{jk}^i$  and is found using a least squares method (Figure 35). The cost for a particular join is computed by summing the squared residuals for each cepstral coefficient:

$$cost(j, k) = \sum_{\forall i} ([l_j^i \ r_k^i] - b_{jk}^i * [1 \ \dots \ 2N])^2$$

<sup>18</sup> Due to the relatively small size of our database we have empirically determined that the value of *perNewUnits* should be no more than 75% (i.e. *perNewUnits*  $\in [0,0.75]$ ).

The optimal cut point is specified by the  $[j \ k]$  pair with the minimum cost value.

One of the disadvantages to optimal coupling is the possibility of significantly altering the final duration of a phone. A possible solution is to constrain the potential join points to those that do not alter the final duration, but Conkie and Isard conclude that this is too restrictive and produces sub-optimal joins (Conkie and Isard, 1997). Our solution incorporates a weighting parameter  $\beta$  that penalizes a join for deviating from the desired target duration. The duration penalty is given by

$$DP(j, k) = 1 + \beta * \max\left(\frac{\text{duration}(j, k)}{\text{targetDuration}} - 1, \frac{\text{targetDuration}}{\text{duration}(j, k)} - 1\right)$$

where  $\text{duration}(j, k)$  is a function that calculates the final duration of the shared phone that results from joining two diphones (Figure 34). The final cost is given by

$$\text{cost}(j, k) = \sum_{\forall i} ([l_j^i \ r_k^i] - b_{jk}^i * [1 \ \dots \ 2N])^2 * DP(j, k)$$

In preliminary tests, we found  $\beta = 0.33$  to provide a good balance between join smoothness and accurate durations.

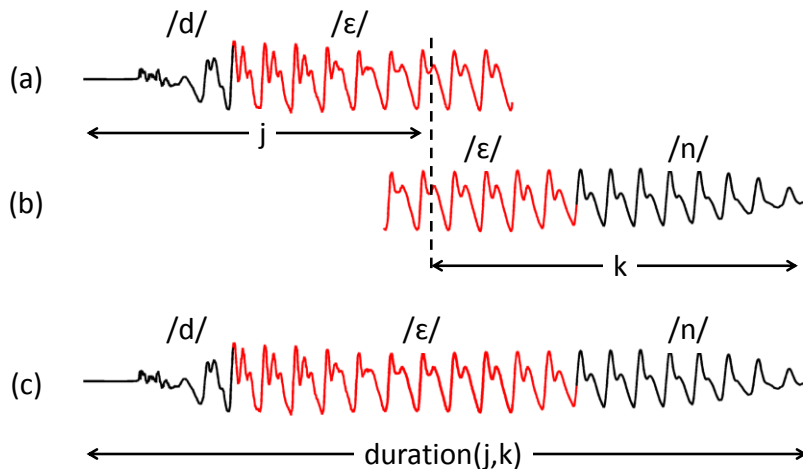


Figure 34 The left diphone (a) and right diphone (b) are joined to form a triphone (c). The join location is specified by the points  $j$  and  $k$ . The duration of the concatenated triphone is  $j+k$ .

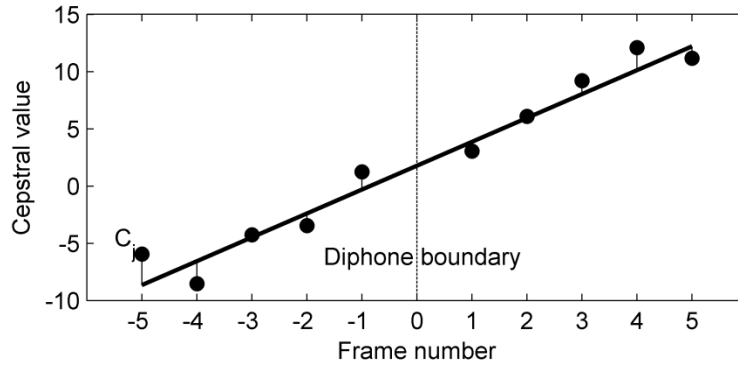


Figure 35 Calculating the cost of joining two diphones. The concatenation cost is given as the sum of squared residuals for each cepstral component  $C_j$ .

### 6.3.3 Spectral smoothing with pulse-density modulation

Optimal coupling offers sufficient smoothing when an acoustically stable join point exists, but its smoothing power is limited because it does not actually modify the spectrum. Spectral interpolation is needed to handle large spectral discontinuities, but linear interpolation does not model the natural movement of formants. Recently, a method based on pulse density modulation (PDM) has been proposed to interpolate STRAIGHT spectral envelopes in a way that shifts formants and amplitudes naturally (Shiga, 2009).

PDM employs a delta-sigma modulator to convert a STRAIGHT spectral envelope  $x(n)$ , where  $n$  denotes frequency, into a pulse sequence  $y(n) = PDM[x(n)]$  as follows:

$$e(n) = x(n) - v_c y(n - 1)$$

$$r(n) = e(n) - r(n - 1)$$

$$y(n) = \text{sign}(r(n))$$

with initial conditions  $r(1) = e(1) = x(1)$  and  $y(n) = 0$ ; the term  $v_c$  represents the feedback gain of the delta-sigma modulator:  $v_c = \max(x)$ .

In turn, the pulse sequence  $y(n)$  can be decoded back into a log spectral envelope  $\hat{x}(n) = PDM^{-1}[y(n)]$  through the discrete cosine transform (DCT) as:

$$c(n) = DCT[y(n)]$$

$$c(n) = 0 \forall n > k$$

$$\hat{x}(n) = DCT^{-1}[c(n)] \times v_c$$

which essentially acts as a low-pass filter by truncating the DCT expansion with an appropriate cutoff  $k$  ( $k = 100$  in our implementation). Thus, given a pair of spectral envelopes  $x_1(n)$  and  $x_2(n)$ , a morphed spectral envelope can be computed by averaging the position of corresponding pulses in the two spectra:

$$x_m(n) = PDM^{-1}[\alpha PDM[x_1(n)] + (1 - \alpha)PDM[x_2(n)]]$$

where the morphing coefficient  $\alpha$  ( $0 \leq \alpha \leq 1$ ) can be used to generate a continuum of morphs between the two spectral envelopes  $x_1(n)$  and  $x_2(n)$ .

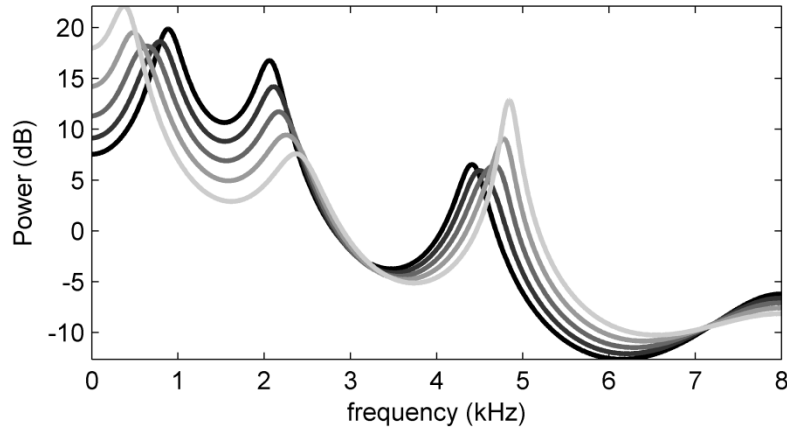


Figure 36 Morphing two spectrums with the PDM method. The original spectrums ( $\alpha = 0,1$ ) are plotted with the darkest and lightest colored lines.

### 6.3.4 STRAIGHT synthesis

STRAIGHT is a high-resolution channel-vocoder. It encodes speech into two source parameters (i.e. fundamental frequency and aperiodicity) and one spectral parameter (i.e. smoothed spectrogram) (Kawahara, 1997). Figure 37 shows the parameters for a portion of MAB speech. ConFAC relies upon this versatile encoding for a variety of applications. Our immediate goal, diphone concatenation, is performed by simply concatenating the parameters associated with each diphone. Spectral smoothing (sub-section 6.3.3) is performed by morphing the smoothed spectrogram around the point of concatenation. MFCCs are calculated by warping the smoothed spectrogram according to the Mel-frequency scale and then computing the discrete cosine transform. Simple voice conversion (sub-section 7.3) is performed by warping the spectral envelope and shifting the fundamental frequency. Furthermore, synthesis from STRAIGHT parameters<sup>19</sup> yields a higher quality of speech than can be obtained through traditional PSOLA methods.

---

<sup>19</sup> Due to the time consuming nature of these calculations, these values are calculated once for the entire database and stored offline. The encoding is highly redundant, requiring more than 128 times the amount of storage than a raw speech waveform (16 kHz). Storage for the MAB and RGO databases requires over 80 GB of storage space.

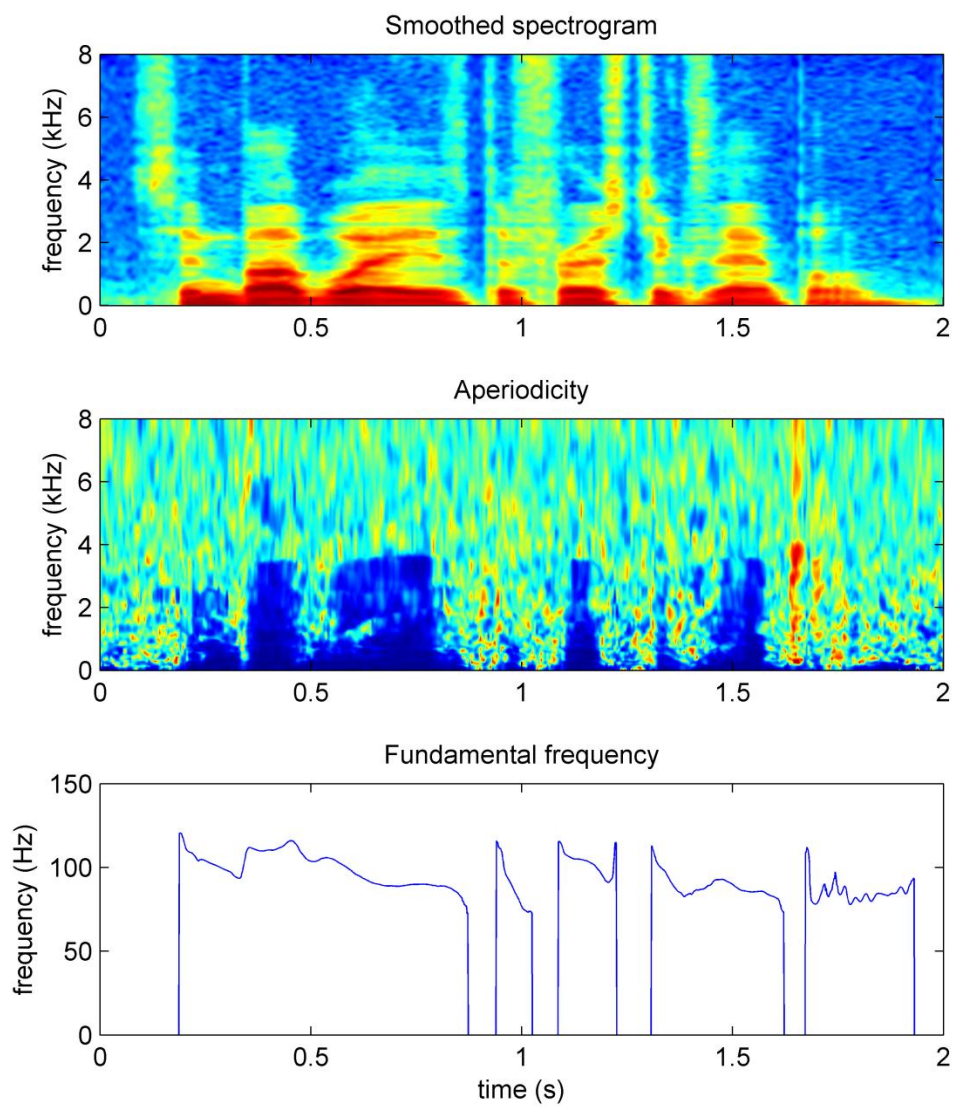


Figure 37 STRAIGHT parameterization of the MAB utterance “Zimbabwe is the first example.”



## **6.4 Conclusion**

Concatenative foreign accent conversion provides a framework to synthesize speech using either articulatory or acoustic features. Accent conversion is performed by selecting diphones from a database of non-native speech that match the articulatory/acoustic patterns of a native speaker. In the next section we evaluate ConFAC's in terms of its ability to transform accent using acoustic and articulatory features.

## 7. CONFAC EXPERIMENTS

Three experiments are performed to evaluate ConFAC. The first two measure ConFAC's ability to alter the accent of a non-native speaker of English. The final experiment examines whether articulatory encodings of speech are more linguistically stable than acoustic representations.

### 7.1 Subject recruitment

Experiments in this section were evaluated by participants through Amazon's online crowdsourcing tool: Mechanical Turk. Mechanical Turk has traditionally been used by companies to perform "Human Intelligence Tasks" that are difficult for computers to perform, such as image tagging or speech transcription. It can also be used to solicit questionnaires or perform user studies (Callison-Burch, 2009; Kittur et al., 2008). It allows experiments to be run in a matter of hours rather than weeks, although there is an initial overhead required to create tests in a web-based format and some experimental designs do not lend themselves to this format. The testing environment of Mechanical Turk is less controlled than the subjective tests from Section 4. In particular, we have no control over listening conditions (e.g. headphones/speakers or television/music in the background), nor is it possible to ensure that participants gave their complete attention to the task at hand. To maximize the number of native speaking participants, a pre-test qualification required potential participants to correctly identify various American accents. Participants who did not pass this qualification were not allowed to participate in the study. In addition, a post-test prompt asked users 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. Examples of the web-based forms are provided in APPENDIX E.

## 7.2 Experiment #1 – Accent rating

The first experiment repeats the subjective accent rating from sub-section 4.3, but in this case for stimuli created using ConFAC. Two experimental conditions were tested: accent conversion in MFCC space<sup>20</sup> ( $AC_{MFCC}$ ) and accent conversion in Maeda space<sup>20</sup> ( $AC_{Maeda}$ ). We tested values for the parameter *perNewUnits* from 0.1 to 1.0 in increments of 0.1. The final value of 0.5 was informally determined to be the highest tested value that did not significantly alter the overall level of naturalness of synthesized utterances. This corresponds to replacing 50% of the diphones in the non-native utterance with different diphones from the non-native corpus. The same 10 sentences (listed in Table 6) were evaluated for original recordings of the foreign and native speakers and the experimental conditions. Twenty participants rated 40 utterances on a 7 point EGWA scale (0=not at all accented, 2= somewhat accented, 4=quite a bit accented, 6=extremely accented).

Table 6  
Transcripts of the 10 sentences used in Experiments 1 and 2 of this section.

Test sentences
The obvious answer is cash.
He said: Education, education, education.
They are so easy for youngsters to open.
There was huge irony here.
It is due for release in the U K early next year.
Everybody meddles with nature.
This is the big fear.
We must take a measured look at this.
We are regarded as being dour people.
This was a meeting which changed his life.

<sup>20</sup> The features pitch, loudness, and phoneme duration were included in all conditions. These are technically acoustic measurements, but they are representative of similar articulatory features: glottal activity (frequency and power) and rate of speech.

### 7.2.1 Results

Results from the first experiment indicate a large difference in perceived accent between RGO and MAB, but  $AC_{MFCC}$ , and  $AC_{Maeda}$  were rated similar to RGO (Figure 38). A repeated measures ANOVA test was performed to test the null hypothesis that the average rating of accent for RGO,  $AC_{MFCC}$ , and  $AC_{Maeda}$  are the same. The results do not give sufficient evidence to reject the null hypothesis, i.e. there is no significant difference in perceived accent  $F(2,38)=0.52, p=0.60$ . We suspect that the high similarity between RGO,  $AC_{MFCC}$ , and  $AC_{Maeda}$  may have influenced listener ratings. Namely, we believe listeners assigned similar ratings to avoid cognitive dissonance. Cognitive dissonance is the unpleasant feeling caused by holding conflicting beliefs (Festinger, 1957). The theory of cognitive dissonance states that people strive to reduce dissonance by changing one or more beliefs. In this study, we believe subjects avoided cognitive dissonance by assigning consistent accent ratings to recordings from the “same speaker” (RGO,  $AC_{MFCC}$ , and  $AC_{Maeda}$  sounded very similar). We tested this conjecture in the following two-part experiment.

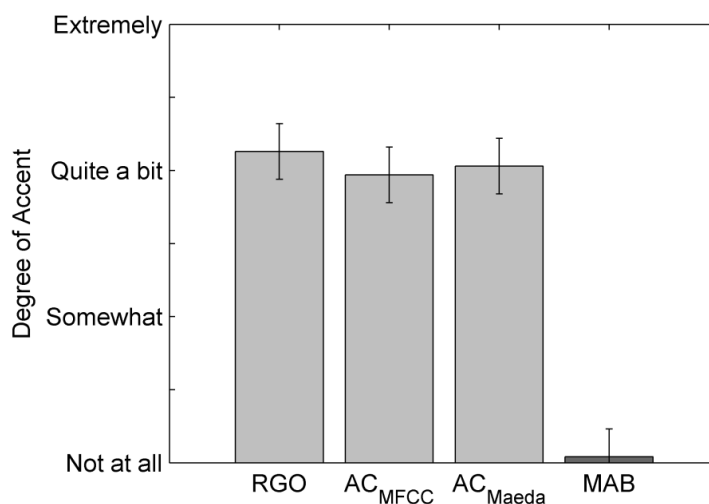


Figure 38 Accent ratings for ConFAC. The error bars indicate intervals of confidence ( $\alpha=0.05$ ) in a multiple comparison test.

### 7.3 Experiment #2- Decoupling accent and identity (Part 1)

This experiment measures the extent to which accent ratings were affected by the perceived identity of a speaker. Namely, we hypothesize that listeners in Experiment #1 resisted assigning ratings that they perceived to be inconsistent (i.e. providing significantly different responses to recordings from the “same” speaker). To permit listeners to assign internally consistent ratings, we wish to disguise the experimental conditions ( $AC_{MFCC}$  and  $AC_{Maeda}$ ) by changing their identity. The first part of the experiment tests whether changing the identity (without accent conversion) affects the perception of accent. For this purpose, we disguise the original RGO and MAB recordings. Three baseline guises were created:

- 1)  $G_{ave}$ : modeled after RGO, this guise resembles an average male voice,
- 2)  $G_{deep}$ : modeled after MAB, this guise resembles a deep male voice, and
- 3)  $G_{child}$ : modeled as the identity opposite to  $G_{deep}$ , this guise resembles a child voice.

Applying a particular guise to a voice consists of altering its fundamental frequency and long-term spectral properties. The fundamental frequency is shifted and scaled within the range of the target guise (e.g. average fundamental frequency for  $G_{ave}$  is 140 Hz). The STRAIGHT smoothed spectrogram is next modified to match the global statistics of the target guise. This is accomplished via dynamic frequency warping VTLN (Lee and Rose, 1998). Namely, we find a frequency warping function to minimize differences between time-aligned spectrograms of the current speaker and target guise. As illustrated in Figure 39, the frequency warping function  $F(w)$  is used to disguise a MAB utterance to the average guise. Conversely, applying the  $G_{deep}$  guise to RGO uses the inverse function  $F^{-1}(w)$ . Six types of stimuli were created for this test by combining two source voices (RGO and MAB) with three guises (Table 7). Twenty participants rated 10 utterances from each condition on a 7 point EGWA scale (0=not at all accented, 2=somewhat accented, 4=quite a bit accented, 6=extremely accented).

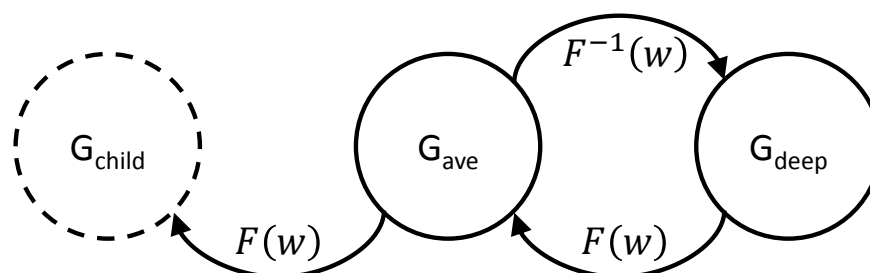


Figure 39 Transforming a speaker to a particular guise. Guises  $G_{\text{ave}}$  and  $G_{\text{deep}}$  are modeled after RGO and MAB, respectively. The baseline warping function  $F(w)$  is the result performing dynamic frequency warping from MAB to RGO.

Table 7  
Six conditions used in part 1 of Experiment #2.

Source voice	Guise	Transformation	Notes
RGO	deep	$F^{-1}(w)$	-
RGO	ave	-	original RGO
RGO	child	$F(w)$	-
MAB	deep	-	original MAB
MAB	ave	$F(w)$	-
MAB	child	$F(F(w))$	-

### 7.3.1 Results

A two-factor repeated measures ANOVA test was performed for the factors “source voice” and “guise.” The results show a significant difference in source voice (i.e. RGO or MAB)  $F(1,97)=1812.08$ ,  $p<0.001$ , but no significant difference for guise  $F(2,97)=0.79$ ,  $p=0.46$ . In other words, modifying the fundamental frequency and long term spectral properties of a voice does not affect its accent rating (Figure 40). This is a positive result because it allows us to apply guises to stimuli in Experiment #1 without affecting their *true* accent ratings.

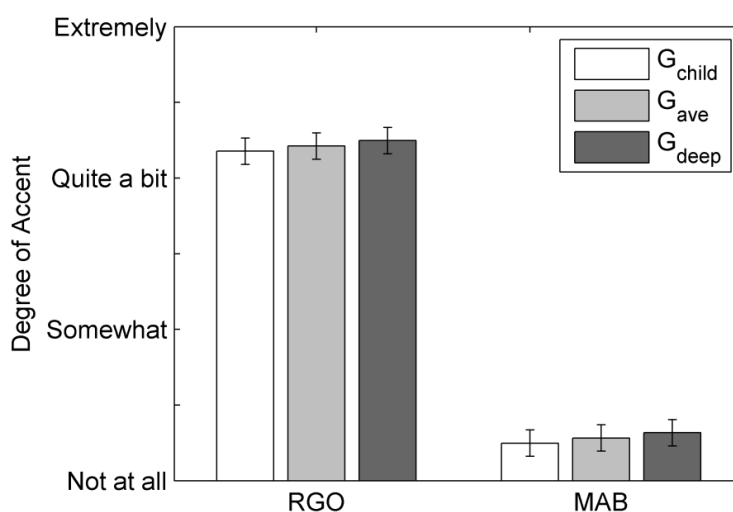


Figure 40 Accent ratings for the change of identity experiment. The error bars indicate intervals of confidence ( $\alpha=0.05$ ) in a multiple comparison test.

### 7.4 Experiment #2- Decoupling accent and identity (Part 2)

The second half of the experiment tests whether listeners in Experiment #1 rated  $AC_{MFCC}$ ,  $AC_{Maeda}$  and RGO as having a similar accent because they perceived them as the same speaker. In this case, we disguised  $AC_{MFCC}$  and  $AC_{Maeda}$  with the previously developed guises. Two separate tests were performed to balance the choice of disguise across experimental

conditions. In the first test (denoted by Test Set A in Table 8),  $AC_{MFCC}$  and  $AC_{Maeda}$  underwent  $G_{child}$  and  $G_{deep}$  transforms respectively. Listeners rated these utterances in addition to unmodified RGO and MAB utterances. Test set B switched the guises of the experimental conditions.

Table 8  
Separation of stimuli into two test sets (A and B).

Set A	Set B	Condition	Guise	ConFAC synthesis features <sup>20</sup>
✓	✓	Foreign (RGO)		-
✓		$AC_{MFCC}$	$G_{child}$	MFCC
✓		$AC_{Maeda}$	$G_{deep}$	Maeda
	✓	$AC_{MFCC}$	$G_{deep}$	MFCC
	✓	$AC_{Maeda}$	$G_{child}$	Maeda
✓	✓	Native (MAB)		-

#### 7.4.1 Results

Results from tests A and B are combined in a repeated measures ANOVA analysis to test the null hypothesis that the means of RGO,  $AC_{MFCC}$ , and  $AC_{Maeda}$  are the same. The evidence suggests that we can reject the null hypothesis; there is a significant difference between the means  $F(2,78)=16.08, p<0.001$ . The results of a multiple comparison test show that all pairs are significantly different except for  $AC_{MFCC}$  and  $AC_{Maeda}$  (Figure 41). In this case, the perceived accent of  $AC_{MFCC}$  and  $AC_{Maeda}$  are 16% and 20% lower than RGO. This result indicates that listeners in Experiment #1 were biased by the similarity of the stimuli. We believe that the guises in this experiment allowed listeners to assign internally consistent ratings by creating perceptually distinct speakers. This technique effectively decouples accent from identity. More importantly, the test showed that both accent conversions can reduce the accent of the original recordings from RGO.



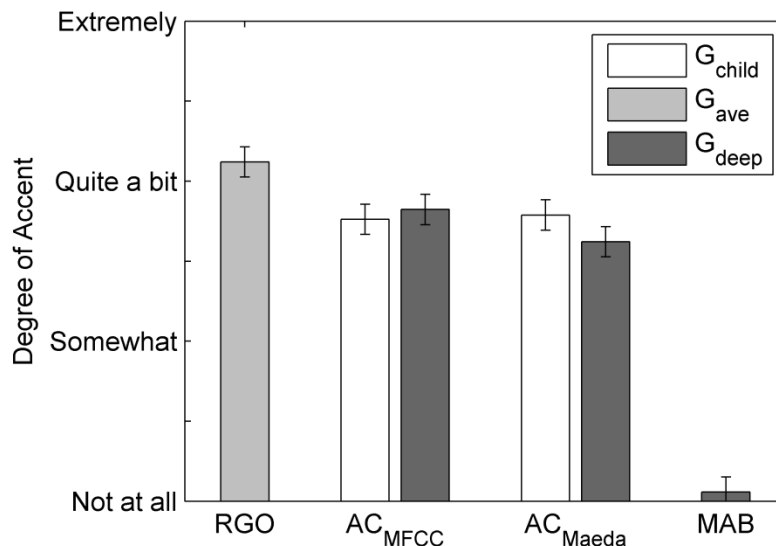


Figure 41 Accent ratings for the experimental conditions after undergoing a change of identity. The error bars indicate intervals of confidence ( $\alpha=0.05$ ) in a multiple comparison test.

### 7.5 Experiment #3 - Comparing features for cross-speaker synthesis

ConFAC's ability to choose more "native sounding" diphones depends primarily on the quality of the synthesis features. The ideal set of synthesis features would contain full linguistic information and be speaker independent. The objective of the third experiment is to compare the articulatory and acoustic feature sets along these two metrics. We compare the linguistic content in each set by performing synthesis *within* speaker. For this purpose, RGO utterances are resynthesized in a leave-one-out fashion. Utterances created in this manner are called same-speaker (SS). We compare speaker dependence of each set by performing synthesis *across* speakers. In such a case, the encoding (MFCC or Maeda) that is more similar across speakers will have an advantage. For this purpose, units from RGO are selected and combined based on their similarity to MAB's synthesis features. Utterances created in this manner are called cross-speaker (CS).

Stimuli in this experiment were created differently from those of the previous experiments. We revert to the original unit selection formulation proposed by Hunt & Black (1996) rather than the adaptations presented in sub-section 6.3.1.1 (allowing original units and modifying the total cost equation). This ensured the selection of new diphones from RGO’s speech corpus for *all* diphones in an utterance and maximized the differences between utterances. However, due to the small size of our speech corpus, the quality of the synthesized utterances is significantly lower than stimuli used in the first two experiments. One-hundred utterances were created by synthesizing 25 unique sentences in each of the four conditions shown in Table 9.

Table 9  
Experimental conditions in the cross-speaker synthesis test.

Condition	Set A	Set B	Synthesis database (speaker)	Synthesis features (speaker)	Target features <sup>20</sup>
$SS_{MFCC}$	✓		RGO	RGO	MFCC
$SS_{Maeda}$	✓		RGO	RGO	Maeda
$CS_{MFCC}$		✓	RGO	MAB	MFCC
$CS_{Maeda}$		✓	RGO	MAB	Maeda

Participants heard the utterances in pairs and were asked to select the more “natural and intelligible” utterance. Utterance pairing was performed as follows: stimulus set A was created by pairing identical sentences (e.g. “They are so easy for youngsters to open.”) from the same-speaker conditions ( $SS_{MFCC}$  and  $SS_{Maeda}$ ) in a random order with a 1 second pause between them. This set allowed us to compare the linguistic information in each feature set since the synthesis features did not contain speaker dependent information from another speaker. Stimulus set B paired the cross-speaker conditions  $CS_{MFCC}$  and  $CS_{Maeda}$  in a similar manner. Cross-speaker utterances are influenced by both linguistic and speaker dependent information. In order to

isolate the amount of speaker dependent information in each set, we measure subject responses relative to the responses in set A. Twenty participants selected the more “natural and intelligible” utterance for each of the 50 stimuli (set A and B).

### 7.5.1 Results

Results from the cross-speaker synthesis test were analyzed with the binomial significance test and McNemar’s (Chi-squared corrected) matched pair test. A two-tailed binomial significance test was performed independently on stimulus set A and B with the null hypothesis that there was an equal preference ( $P=0.5$ ) for the choice of target features (i.e. MFCC and Maeda). Out of the 500 responses in set A (20 participants  $\times$  25 stimuli), 321 favored MFCC compared to 179 for Maeda. This corresponds to a preferred Maeda proportion of only 0.358, which is statistically significant  $p(\text{two-tailed}) < 0.001$ . We attribute the preference for  $SS_{\text{MFCC}}$  to the known shortcomings of the Maeda parameters: 1) they provide an incomplete representation of the vocal tract, 2) sensor values drift over time, and 3) small changes in the articulatory space can produce large changes in speech (e.g. /p/, /t/, /b/, /d/, /k/). Responses for set B, on the other hand, did not provide significant evidence to reject the null hypothesis with 260 responses favoring Maeda compared to 240 for MFCC. In summary, there is a strong preference for acoustic-driven synthesis when the synthesis features are selected from the same speaker as the synthesis database, but there is no clear preference for cross speaker synthesis.

We next employ McNemar’s test to see if this difference is statistically significant.

McNemar’s matched pair test combines listener responses in set A and B using a  $2 \times 2$  contingency table as follows. For each of the 25 unique sentences chosen for this test, a subject’s A and B responses are paired to index one of the four central bins in Table 10. For example, if a listener preferred  $SS_{\text{MFCC}}$  (over  $SS_{\text{Maeda}}$ ) for “sentence 1” and  $CS_{\text{Maeda}}$  (over  $CS_{\text{MFCC}}$ ) for the same sentence then that response is recorded in the upper-right bin. The null hypothesis for

McNemar's matched pair test states that the row and column marginal frequencies are equal for each outcome. In other words, due to the high preference for MFCCs in the same-speaker condition we should expect a similar preference in the cross-speaker condition. Our results show strong evidence to reject the null hypothesis; the Maeda preference proportion increased from 0.358 in the same-speaker condition to 0.52 in the cross-speaker condition ( $\chi^2 = 26.7782$ ,  $p(\text{two-tailed}) < 0.001$ ).

Table 10  
Contingency table showing the results of the pre-test.

		cross-speaker		row total
		MFCC	Maeda	
same-speaker	MFCC	161	160	321
	Maeda	79	100	179
column total		240	260	500

This result supports our hypothesis that speech is more universally represented in the articulatory domain than the acoustic domain for this pair of speakers. The conclusion is drawn from the relative improvement (i.e. 0.358 to 0.52) and not the final value of the cross-speaker test (0.52). This is analogous to a race in which one runner was given a head start and the final result was a tie. The experimental design used here is quite powerful, but it cannot pinpoint the root cause of the preference shift. Subjects displayed an increased preference for Maeda in the cross-speaker condition, but this shift has three possible explanations: increased preference for Maeda, decreased preference for MFCC, or some combination of the two. Fortunately, all explanations boil down to a single conclusion: Maeda parameters were more similar than MFCCs for the tested speakers. Additional studies are required to confirm that this result holds true for more speaker combinations.

This finding is important because it addresses a common issue among investigations of articulatory similarity across speakers. Westbury et al. (1998) declare that “*differences between talkers have been found in every speech kinematic study with a sample greater than one.*” Investigations comparing articulatory gestures across speakers commonly observe articulatory differences that cannot be explained by physical factors (e.g. vocal tract length, palate shape) or normalization processes (Hashi et al., 1998; Johnson et al., 1993; McGowan and Cushing, 1999; Simpson, 2002; Toth and Black, 2005; Westbury et al., 1998). These results are often used to weigh in on the divisive topic of whether or not speakers aim for auditory or articulatory targets. However, such studies consistently overlook a critical factor—the acoustic consequences of articulatory differences. It is well known that the articulatory-to-acoustic mapping is many-to-one (i.e. multiple vocal tract configurations can produce similar acoustic spectrums). In other words, articulatory differences do not necessarily imply acoustic differences. Furthermore, certain acoustic differences may not change the perception of a phoneme (e.g. coarticulation causes phones to sound different depending on their context, but they are still *perceived* as the same phoneme). ConFAC’s ability to estimate<sup>21</sup> the acoustic consequences of articulatory differences makes the results obtained in Experiment #3 unique.

## 7.6 Discussion

ConFAC reduced the accent of a non-native speaker up to 20%. This is considerably less than the results obtained for SpFAC, which reduced accent by 60%. I believe the reason lies in SpFAC’s direct use of the native speaker’s speech whereas ConFAC was limited to the inventory of speech segments from the non-native speaker’s speech corpus. ConFAC assumes that the non-

---

<sup>21</sup> Unlike a traditional articulatory synthesizer, ConFAC cannot produce sounds that are not contained in the database; it can only produce the sound of the nearest articulatory configuration.

native corpus contains diphones with different degrees of accent and it uses features from a native speaker to locate the least accented examples. However, if the non-native speaker is consistent in their production throughout the corpus, then it is futile to make any replacements. ConFAC is also more likely to be influenced by the choice of non-native speaker than SpFAC. The L1/L2 language pairing, particular mispronunciations, and the size and quality of the non-native speech corpus affect ConFAC's ability to transform accent. Unfortunately, due to the rarity of speech corpora with simultaneous acoustic and articulatory recordings, we were only able to experiment with a single non-native speaker.

The RGO corpus contained about half of the number of phones recommended to perform diphone synthesis ( $20,000 < 36,000$ ) (Clark et al., 2007). Dr. Tracy Hammond suggested a way to effectively increase the number of available units in the database without performing any additional data collection. She suggested to create variations of each unit by applying simple prosodic modifications (i.e. pitch, loudness, duration). This approach is only valid insofar as the modifications do not affect the underlying articulatory parameters. For example, the faster a phone is pronounced the more heavily the articulators are influenced by coarticulation. Although we can alter the duration of a phone with a high-quality prosodic modification, we cannot predict how the underlying articulatory gestures should change accordingly. Another way to address this problem is to determine the potential benefit of a larger database. How do the size and coverage of the non-native database affect ConFAC's ability to transform accent? In experiments #1 and #2 we replaced half the diphones with new ones from the non-native corpus. A larger database might allow the replacement of all diphones, but I do not expect twice the reduction in accent due to the tendency of unit selection to first replace those units with the highest target costs. Therefore diphones that are replaced should be among

the most “mispronounced” and the replacement of additional units should result in diminishing improvements.

We did not formally evaluate ConFAC using subjective identity or quality measures. However, strong evidence that the identity of the non-native speaker was maintained is provided by the fact that we needed to disguise stimuli to observe a measurable difference in listener ratings of accent. With regard to quality, we noticed that synthesis quality varied from sentence to sentence. This is most likely caused by the small size of our database. The stimuli used in Experiments 1 and 2 of this section were partially selected based on their quality and naturalness compared to the unmodified RGO recordings.

In light of this section’s results, is it possible to glean any new insight from previous results? SpFAC’s segmental transform reduced accent by 60%, but this result was obtained without resorting to guises. According to the identity test with forward speech, listeners already perceived it as a third speaker. There are two reasons for SpFAC’s segmental transform to sound like a third person: 1) VTLN did not remove all information pertaining to the native speaker and 2) the change of accent was significant enough to act as its own guise. Another insight can be found by revisiting results for the prosodic transform. SpFAC’s prosodic transform did not significantly affect accent ratings, but it was not differentiated in the perceptual identity space. It may have a measurable effect if we were to repeat the experiment with a guise for prosodically modified stimuli.

## 8. CONCLUSION

Speech modification algorithms have been used to alter pitch, duration, and even the identity of speech. Recently, researchers have turned to the problem of modifying accent (Huckvale and Yanagisawa, 2007) (Yan et al., 2007) (Yanguas et al., 1999). Previous approaches rely upon acoustic cues to transform speech. We predicted that the task might be better suited in the articulatory domain. To test this hypothesis we collected a custom speech corpus using an electromagnetic articulograph and developed a speech synthesis system that could be controlled by either acoustic or articulatory features.

Concatenative foreign accent conversion (ConFAC) combined segments of speech from a non-native corpus to maximize the acoustic/articulatory similarity with a native utterance. The approach is particularly appealing because it modifies speech in a way that produces realistically achievable changes in accent (important for applications in pronunciation training). Using this approach, we showed that the degree of foreign accent in a Spanish speaker of English can be reduced by 20%. The experiment performed in sub-section 7.5 confirmed that articulatory-based features were more speaker-independent than an acoustic encoding, but it also indicated that the articulatory encoding was incomplete (i.e. there was insufficient information to reconstruct the speech signal). This is most likely due to the fact that the electromagnetic articulograph measures only 6 points within the vocal tract (upper and lower lips, jaw, and 3 points along the tongue), whereas acoustic features characterize the full vocal tract. We conclude that both representations are equally suitable for accent conversion.



A second contribution of this research was the development of subjective and objective measures to assess the performance of accent conversion systems. The measures were thoroughly tested using spectral foreign accent conversion (SpFAC). SpFAC achieved 60% reduction in accent, but the transformation also altered the perceived identity of the stimulus. In order to overcome difficulties inherent to the evaluation of accent conversion<sup>22</sup>, we proposed two noteworthy experimental designs: 1) a measure of identity that used reverse speech to remove a speaker's accent, and 2) a measure of accent that used voice disguises to eliminate bias caused by similar identities.

The relationship between accent and identity made it difficult to evaluate accent conversion, but it resulted in findings that may be generally useful to speech perception. Our work yielded two insights into the relationship between accent and identity:

- **Large changes in accent affect identity.** Results from the *forward speech* identity test show that listeners perceive stimuli undergoing SpFAC's segmental transform as neither the foreign nor native speaker. We believe this to be partially<sup>23</sup> caused by a simultaneous change in perceived accent (60%). Evidence to support this claim is found in the *reverse speech* identity test where results confirm the voice quality of the foreign speaker is maintained (Figure 21).
- **Small changes in accent are masked by identity.** Results from the (undisguised) accent test show that listeners rate ConFAC stimuli and unmodified foreign stimuli equally, regardless of any changes introduced by ConFAC (Figure 38). Since VTLN

---

<sup>22</sup> In most cases, success must be measured without the benefit of having a well-defined goal. However, a ground truth would exist in cases involving bidialectal speakers or identical twins that each speak a different dialect.

<sup>23</sup> We also suspect that VTLN did not remove all speaker dependent information from the native speaker's spectral envelope.

guises alone do not affect accent, the reduced accent observed in the disguised accent test must be attributed to ConFAC (Figure 41). Therefore, the same reduction of accent may have been present in the first test but not detected.

Figure 42 presents a hypothetical accent conversion continuum consistent with the observed perceptual interactions between accent and identity. The continuum progresses from RGO to ConFAC to SpFAC. The end of the continuum is labeled by a question mark to denote the fact that the identity is not midway between foreign and native identities, but rather an unknown third speaker. We also predict that this effect happens non-linearly near the point where the accent is different enough from RGO that naïve listeners perceive him as a third speaker.

These observations call for a re-evaluation of the goals of accent conversion. Up to this point, we have loosely used the term “identity” to refer both to a speaker’s “voice quality” and their overall persona. But it is possible to change a speaker’s identity yet maintain their voice quality. Voice quality is just one of the many factors affecting our perception of identity. The goal of accent conversion should be to alter a speaker’s accent and maintain *voice quality*. Under this new definition, SpFAC can be considered one step closer towards that goal.

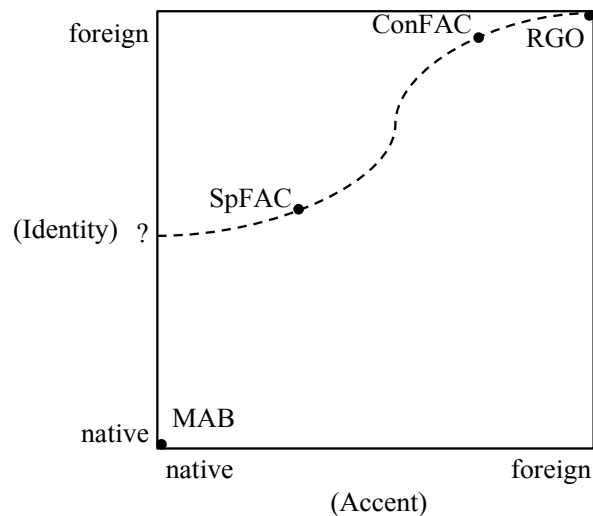


Figure 42 Graphical representation of the relationship between accent and identity. The dashed line represents a hypothetical accent conversion continuum. The location of SpFAC is also hypothetical because SpFAC was evaluated for the speakers KSP and RMS from the ARCTIC database (Kominek and Black, 2003).

### 8.1 Future work

Computer assisted pronunciation training is an exciting application for accent conversion. We believe that accent conversion may provide second language learners with a type of feedback that is both engaging and useful—a more native sounding version of their own speech. SpFAC or ConFAC could potentially be used in a pronunciation training environment to test this claim. Based on our results, SpFAC is the more likely candidate since it achieves a lower accent, with less processing, and only a small amount of acoustic training data. On the other hand, ConFAC may be more encouraging for L2 speakers because it offers a more attainable goal (ConFAC merely resequences previously pronounced sounds). It may also be desirable to post process utterances with a head-transfer function to approximate how users hear their own voice (Huopaniemi et al., 1999).

A possible criticism of this work is the small number of speakers tested. The primary reason for this is the scarcity of articulatory speech corpora. We required datasets that included a

large number of utterances for speakers with distinct foreign accents. ConFAC had the additional constraint that the database must also include articulatory information. No suitable databases existed, so we collected our own. A secondary reason for the limited number of results is the effort associated with evaluation. The objective measures from section 5 might be suitable for tuning parameters, but the final evaluation should always rely on human perception. Formal results for SpFAC were presented for the speaker pairing consisting of the native speaker *rms\_usmale2* and non-native speaker *ksp\_indianmale* from the ARCTIC speech corpus (Kominek and Black, 2003). Two other speakers were considered including native speaker *bdl\_usmale1* and *awb\_scottishmale*. We selected the final pairing from four possible pairings based on informal evaluations. In this regard, SpFAC had an advantage over ConFAC because the latter had no alternatives but to use RGO and MAB. Future studies with different speakers and accents may demonstrate improved ConFAC performance.

We believe ConFAC was also partially limited by the approach adopted for weight training. Unit selection synthesis requires weights for each feature to determine their relative contribution to the target cost (sub-section 6.3.1). ConFAC uses the standard regression training approach proposed by Hunt and Black (1996), which finds weights to minimize the difference between a natural utterance and the output of the synthesizer given the synthesis features of the natural utterance. Since this approach relies only upon the database of the synthesis speaker (RGO), it may not be optimal when units are selected using features from MAB<sup>24</sup>. One solution would be to train weights on a population of speakers so that the resulting weights are generally reliable. However, this was not feasible here since it requires a large amount of articulatory data

---

<sup>24</sup> Accent conversion uses MAB features for the fraction of diphones determined by the synthesis parameter *perNewUnits* (default 50%).

for numerous speakers. Our current solution is to use the weights trained on RGO's database and carefully normalize MAB features to be within the expected range.

A natural extension of ConFAC is to try to combine the benefits of acoustic data (complete information) with those of articulatory data (linguistically relevant). In the current work, two weight training sessions are used: one for  $AC_{MFCC}$  and one for  $AC_{Maeda}$ . To test whether the unified set of features (i.e., articulatory *and* acoustic) could be used for accent conversion, we trained weights for: 13 MFCCs, 5 Maeda parameters, pitch, duration, and loudness. An inspection of the resulting weights reveals that nearly all high weight values belong to MFCCs and utterances synthesized with the hybrid weights were not perceptually different than those obtained for  $AC_{MFCC}$ . We believe this result can be traced back to the limitations of the current weight training procedure (discussed above). Since weights are trained using units and features from a single speaker (RGO), they are optimized for same-speaker synthesis. Results from the same-speaker condition in Experiment #3 show a clear preference for the MFCC condition ( $SS_{MFCC}$ ), which explains the high weight values for MFCC in the hybrid set. Therefore, we do not expect hybrid weights to provide significantly better results until the cross-speaker weight training problem is solved. These are a few of the potential directions for this work.

## REFERENCES

- Abe, M., Nakamura, S., Shikano, K., Kuwabara, H., 1988. Voice conversion through vector quantization, In: Proc. ICASSP, New York, NY pp. 655-658.
- Al Bawab, Z., Bhiksha, R., Stern, R.M., 2008. Analysis-by-synthesis features for speech recognition, In: Proc. ICASSP pp. 4185-4188.
- Anisfeld, M., Bogo, N., Lambert, W.E., 1962. Evaluational reactions to accented English speech. *Journal of Abnormal and Social Psychology* 65, 223-231.
- Arslan, L.M., 1999. Speaker transformation algorithm using segmental codebooks (STASC). *Speech Communication* 28, 211-226.
- Arslan, L.M., Hansen, J.H.L., 1996. Language accent classification in American English. *Speech Communication* 18, 353-367.
- Arslan, L.M., Hansen, J.H.L., 1997. A study of temporal features and frequency characteristics in American English foreign accent. *The Journal of the Acoustical Society of America* 102, 28-40.
- Arthur, B., Farrar, D., Bradford, G., 1974. Evaluation reactions of college students to dialect differences in the English of Mexican-Americans. *Language and Speech* 17, 255-270.
- Barry, W., Hoequist, C., Nolan, F., 1989. An approach to the problem of regional accent in automatic speech recognition. *Computer Speech and Language* 3, 355-366.
- Baudoin, G., Stylianou, Y., 1996. On the transformation of the speech spectrum for voice conversion, In: Proc. ICSLP, pp. 1405-1408.
- Bishop, C.M., 2006. *Pattern recognition and machine learning*. Springer, New York.
- Bissiri, M.P., Pfitzinger, H.R., Tillmann, H.G., 2006. Lexical stress training of German compounds for Italian speakers by means of resynthesis and emphasis, In: Proc. Australian International Conference on Speech Science & Technology, pp. 24-29.
- Boersma, P., Weenink, D., 2007. Praat: Doing phonetics by computer. [www.fon.hum.uva.nl/praat/](http://www.fon.hum.uva.nl/praat/).
- Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 27, 113-120.
- Brennan, E., Brennan, J., 1981. Accent scaling and language attitudes: Reactions to Mexican American English speech. *Language and Speech* 24, 207-221.

- Callison-Burch, C., 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's mechanical turk, In: Proc. Empirical Methods in Natural Language Processing, Singapore, pp. 286-295.
- Campbell, N., 1998. Foreign-language speech synthesis, In: Proc. Workshop on Speech Synthesis, Jenolan Caves, Australia, pp. 117-180.
- Celce-Murcia, M., Brinton, D., Goodwin, J.M., 1996. Teaching pronunciation: A reference for teachers of English to speakers of other languages. Cambridge University Press, Cambridge ; New York.
- Chassot, M., 2009. Magnetic Field Optimization from Limited Data. Master Thesis. University of Edinburg.
- Chen, T., Huang, C., Chang, E., Wang, J., 2001. Automatic accent identification using Gaussian mixture models, In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, Madonna di Campiglio, Italy, pp. 343-346.
- Chun, D.M., 1998. Signal analysis software for teaching discourse intonation. Language Learning & Technology 2, 61-77.
- Chung-Hsien, W., Chi-Chun, H., Te-Hsien, L., Jhing-Fa, W., 2006. Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis. IEEE Transactions on Audio, Speech and Language Processing 14, 1109-1116.
- Clark, R.A.J., Richmond, K., King, S., 2007. Multisyn: Open-domain unit selection for the Festival speech synthesis system. Speech Communication 49, 317-330.
- Coe, N., 2001. Speakers of Spanish and Catalan, Learner English. Cambirdge Publishers, pp. 90-112.
- Compton, A.J., 1963. Effects of filtering and vocal duration upon identification of speakers, aurally. Journal of the Acoustical Society of America 35, 1748-1755.
- Conkie, A., Isard, S., 1997. Optimal coupling of diphones, In: Van Santen, J.P.H. (Ed.), Progress in speech synthesis. Springer, pp. 293-304.
- De La Zerda, N., Hopper, R., 1979. Employment interviewers' reactions to Mexican American speech. Communication Monographs 46, 126 - 134.
- Deller, J.R., Hansen, J.H.L., Proakis, J., 2000. Discrete Time Processing of Speech Signals. Wiley-IEEE Press, New York.
- Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39, 1-38.
- Derwing, T., Munro, M., Wiebe, G., 1998. Evidence in favor of a broad framework for pronunciation instruction. Language Learning 48, 393-410.

- Deshpande, S., Chikkerur, S., Govindaraju, V., 2005. Accent classification in speech, In: Proc. IEEE Workshop on Automatic Identification Advanced Technologies, Buffalo, New York, pp. 139-143.
- Dijkstra, E.W., 1959. A note on two problems in connection with graphs. *Numerische Mathematik* 1, 269-271.
- Doddington, G.R., 1985. Speaker recognition: Identifying people by their voices. *Proceedings of the IEEE* 73, 1651-1664.
- Doh-Suk, K., Tarraf, A., 2004. Perceptual model for non-intrusive speech quality assessment, In: Proc. ICASSP, pp. 1060-1063.
- 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, Paris.
- Felps, D., Bortfeld, H., Gutierrez-Osuna, R., 2009. Foreign accent conversion in computer assisted pronunciation training. *Speech Communication* 51, 920-932.
- Felps, D., Geng, C., Berger, M., Richmond, K., Gutierrez-Osuna, R., 2010. Relying on critical articulators to estimate vocal tract spectra in an articulatory-acoustic database, In: Proc. Interspeech, Makuhari, Japan, pp. 1990-1993.
- Felps, D., Gutierrez-Osuna, R., 2010. Developing objective measures of foreign-accent conversion. *Audio, Speech, and Language Processing* 18, 1030-1040.
- Festinger, L., 1957. *A Theory of Cognitive Dissonance*. Stanford University Press.
- Fielding, G., Evered, C., 1980. The Influence of Patients' Speech upon Doctors: The Diagnostic Interview, In: Robert N. St. Clair, H.G. (Ed.), *The Social and Psychological Contexts of Language*. Lawrence Erlbaum Associates, pp. 51-72.
- Flanagan, J.L., 1972. *Speech Analysis, Synthesis and Perception*. Springer-Verlag, Berlin.
- Forney, G.D., Jr., 1973. The viterbi algorithm. *Proceedings of the IEEE* 61, 268-278.
- Geladi, P., Kowalski, B., 1986. Partial least-squares regression: A tutorial. *Analytica Chimica Acta* 185, 1-17.
- Giles, H., Powesland, P., 1975. *Speech style and social evaluation*. Academic Press London.
- Gold, B., Morgan, N., 2000. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. Wiley, New York.
- Gray, P., Hollier, M.P., Massara, R.E., 2000. Non-intrusive speech-quality assessment using vocal-tract models. *IEE Proc. Vision, Image, and Signal Processing* 147, 493-501.



- Griffin, D., Lim, J., 1984. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32, 236-243.
- Hall, J.L., 2000. Application of multidimensional scaling to subjective evaluation of coded speech, In: *Proc. Speech Coding*, pp. 20-22.
- Hansen, T.K., 2006. Computer assisted pronunciation training: The four 'K's of feedback, In: *Proc. International Conference on Multimedia Information Communication Technologies in Education*, Seville, Spain, pp. 342-346.
- Hashi, M., Westbury, J.R., Honda, K., 1998. Vowel posture normalization. *The Journal of the Acoustical Society of America* 104, 2426-2437.
- Hermansky, H., Broad, D.J., 1989. The effective second formant F2' and the vocal tract front-cavity, In: *Proc. ICASSP*, pp. 480-483.
- Hiroya, S., Honda, M., 2004. Estimation of articulatory movements from speech acoustics using an HMM-based speech production model. *IEEE Transactions on Speech and Audio Processing* 12, 175-185.
- Honikman, B., 1964. Articulatory settings, In: Abercrombie, Fry, MacCarthy, Scott, Trim (Eds.), *In honour of Daniel Jones: Papers contributed on the occasion of his eightieth birthday*, pp. 73-84.
- Hoole, P., Zierdt, A., Geng, C., 2003. Beyond 2D in articulatory data acquisition and analysis, In: *Proc. International Conference of Phonetic Sciences XV*, Barcelona, Spain, pp. 265-268.
- Huckvale, M., 2007. ACCDIST: An accent similarity metric for accent recognition and diagnosis, In: Müller, C. (Ed.), *Speaker Classification II*. Springer Berlin / Heidelberg, pp. 258-275.
- Huckvale, M., Yanagisawa, K., 2007. Spoken language conversion with accent morphing, In: *Proc. ISCA Workshop on Speech Synthesis*, Bonn, Germany, pp. 64-70.
- Hui, Y., Young, S., 2006. Quality-enhanced voice morphing using maximum likelihood transformations. *IEEE Transactions on Audio, Speech and Language Processing* 14, 1301-1312.
- Hunt, A.J., Black, A.W., 1996. Unit selection in a concatenative speech synthesis system using a large speech database, In: *Proc. ICASSP*, pp. 373-376.
- Huopaniemi, J., Zacharov, N., Karjalainen, M.G., 1999. Objective and subjective evaluation of head-related transfer function filter design. *Journal of the Audio Engineering Society* 47, 218-239.
- Ikeno, A., Hansen, J.H.L., 2006. The role of prosody in the perception of US native English accents, In: *Proc. Interspeech*, Pittsburgh, PA, USA, pp. 1437-1440.

- ITU-T, 1996. P. 800: Methods for subjective determination of transmission quality. Report.
- Jilka, M., Möhler, G., 1998. Intonational foreign accent: Speech technology and foreign language teaching, In: Proc. ESCA Workshop on Speech Technology on Language Learning, Stockholm, Sweden, pp. 115-118.
- Jin, L., Kubichek, R., 1994. Output-based objective speech quality, In: Proc. IEEE Vehicular Technology Conference, Stockholm, Sweden, pp. 1719-1772.
- Johnson, K., Ladefoged, P., Lindau, M., 1993. Individual differences in vowel production. *The Journal of the Acoustical Society of America* 94, 701-714.
- Kaburagi, T., Honda, M., 1998. Determination of the vocal tract spectrum from the articulatory movements based on the search of an articulatory-acoustic database. *ICSLP*, 433-436.
- Kain, A., 2001. High Resolution Voice Transformation. Ph.D. dissertation. Oregon Health & Science University.
- Kain, A., Macon, M.W., 1998. Spectral voice conversion for text-to-speech synthesis, In: Proc. ICASSP, pp. 285-288.
- Kalin, R., Rayko, D., 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: Proc. ICASSP, pp. 1303-1306
- Kenny, O.P., Nelson, D.J., Bodenschatz, J.S., McMonagle, H.A., 1998. Separation of non-spontaneous and spontaneous speech, In: Proc. ICASSP, pp. 573-576.
- Kittur, A., Chi, E.H., Suh, B., 2008. Crowdsourcing user studies with mechanical turk, In: Proc. Special Interest Group in Computer Human Interaction, Florence, Italy, pp. 453-456.
- Klatt, D.H., 1980. Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America* 67, 971-995.
- Klatt, D.H., 1987. Review of text-to-speech conversion for English. *The Journal of the Acoustical Society of America* 82, 737-793.
- Kominek, J., Black, A., 2003. CMU ARCTIC databases for speech synthesis. Carnegie Mellon University Language Technologies Institute Report. Technical Report CMU-LTI-03-177.
- Kounoudes, A., Naylor, P.A., Brookes, M., 2002. The DYPSA algorithm for estimation of glottal closure instants in voiced speech, In: Proc. ICASSP, Orlando, FL, USA, pp. 349-352.
- Kreiman, J., Papcun, G., 1991. Comparing discrimination and recognition of unfamiliar voices. *Speech Communication* 10, 265-275.

- Kroos, C., Sock, R., Fuchs, S., Laprie, Y., 2008. Measurement accuracy in 3D electromagnetic articulography, In: Proc. International Seminar on Speech Production, Strasbourg, France, pp. 61–64.
- Kruskal, J., 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29, 1-27.
- Kuhn, G.M., 1975. On the front cavity resonance and its possible role in speech perception *The Journal of the Acoustical Society of America* 58, 428-433
- Künzel, H., 2000. Effects of voice disguise on speaking fundamental frequency. *Forensic Linguistics* 7, 199-289.
- Kuwabara, H., Takagi, T., 1991. Acoustic parameters of voice individuality and voice-quality control by analysis-synthesis method. *Speech Communication* 10, 491-495.
- Lambert, W., Hodgson, R., Gardner, R., Fillenbaum, S., 1960. Evaluational reactions to spoken languages. *Journal of Abnormal and Social Psychology* 60, 44-51.
- Lavner, Y., Gath, I., Rosenhouse, J., 2000. The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Communication* 30, 9-26.
- Lee, L., Rose, R., 1998. A frequency warping approach to speaker normalization. *IEEE Transactions on Speech and Audio Processing* 6, 49-60.
- Ling, Z.-H., Richmond, K., Yamagishi, J., Wang, R.-H., 2009. Integrating articulatory features into HMM-based parametric speech synthesis. *Audio, Speech, and Language Processing* 17, 1171-1185.
- Lippi-Green, R., 1997. *English with an Accent: Language, Ideology, and Discrimination in the United States*. Routledge.
- Little, R., Rubin, D., 1987. *Statistical analysis with missing data*. Wiley.
- Loots, L., Niesler, T., 2011. Automatic conversion between pronunciations of different English accents. *Speech Communication* 53, 75-84.
- Maeda, S., 1990. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model, In: Hardcastle, W.J., Marchal, A. (Eds.), *Speech Production and Speech Modelling*. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 131-149.
- Magen, H.S., 1998. The perception of foreign-accented speech. *Journal of Phonetics* 26, 381-400.
- Makhoul, J., Berouti, M., 1979. High-frequency regeneration in speech coding systems, In: Proc. ICASSP, Washington, District of Columbia, USA, pp. 428-431.

- Malayath, N., Hermansky, H., Kain, A., 1997. Towards decomposing the sources of variability in speech, In: Proc. Eurospeech, pp. 497-500.
- Malfait, L., Berger, J., Kastner, M., 2006. P.563-The ITU-T Standard for single-ended speech quality assessment. *IEEE Transactions on Audio, Speech and Language Processing* 14, 1924-1934.
- Martin, P., 2004. WinPitch LTL II: A Multimodal Pronunciation Software. [http://www.winpitch.com/winpitch\\_ltl.htm](http://www.winpitch.com/winpitch_ltl.htm).
- Matsumoto, H., Hiki, S., Sone, T., Nimura, T., 1973. Multidimensional representation of personal quality of vowels and its acoustical correlates. *IEEE Transactions on Audio Electroacoustics* 21, 428-436.
- McGowan, R.S., Cushing, S., 1999. Vocal tract normalization for midsagittal articulatory recovery with analysis-by-synthesis. *The Journal of the Acoustical Society of America* 106, 1090-1105.
- Meier, P., Muller, S., 2009. IDEA: International Dialects of English Archive. <http://web.ku.edu/~idea/index.htm>.
- Minematsu, N., Nakagawa, S., 2000. Visualization of pronunciation habits based upon abstract representation of acoustic observations, In: Proc. Integration of Speech Technology into Learning, London, England, pp. 130-137.
- Moulines, E., Charpentier, F., 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication* 9, 453-467.
- Moulines, E., Laroche, J., 1995. Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication* 16, 175-205.
- Munro, M., 1995. Non-segmental factors in foreign accent: Ratings of filtered speech. *Studies in Second Language Acquisition* 17, 17-34.
- Munro, M., Derwing, T., 1994. Evaluations of foreign accent in extemporaneous and read material. *Language Testing* 11, 253-266.
- Munro, M., Derwing, T., 1995. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning & Technology* 45, 73-97.
- Munro, M., Derwing, T., Burgess, C., 2003. The detection of foreign accent in backwards speech, In: Proc. International Congress of Phonetic Sciences, Barcelona, Spain, pp. 535-538.
- Nagano, K., Ozawa, K., 1990. English speech training using voice conversion, In: Proc. ICSLP, Kobe, Japan, pp. 1169-1172.

- Nakamura, K., Toda, T., Nankaku, Y., Tokuda, K., 2006. On the use of phonetic information for mapping from articulatory movements to vocal tract spectrum, In: Proc. ICASSP, Toulouse, France, pp. 93-96.
- Narendranath, M., Murthy, H.A., Rajendran, S., Yegnanarayana, B., 1995. Transformation of formants for voice conversion using artificial neural networks. *Speech Communication* 16, 207-216.
- Nava, E., Tepperman, J., Goldstein, L., Zubizarreta, M.L., Narayanan, S., 2009. Connecting rhythm and prominence in automatic ESL pronunciation scoring, In: Proc. Interspeech, Brighton, UK, pp. 684-687.
- Neri, A., Cucchiaroni, C., Strik, H., 2002. Feedback in computer assisted pronunciation training: Technology push or demand pull? , In: Proc. CALL conference on CALL professionals and the future of CALL research, Antwerp, Belgium, pp. 179-188.
- Paul, D., 1981. The spectral envelope estimation vocoder. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 29, 786-794.
- Pelham, B., Blanton, H., 2007. *Conducting Research in Psychology, Measuring the Weight of Smoke*, 3rd ed. Thomson Higher Education, Belmont, CA.
- Probst, K., Ke, Y., Eskenazi, M., 2002. Enhancing foreign language tutors - In search of the golden speaker. *Speech Communication* 37, 161-173.
- Purnell, T., Idsardi, W., Baugh, J., 1999. Perceptual and phonetic experiments on American English dialect identification. *Journal of Language and Social Psychology* 18, 10-30.
- Qin, C., Carreira-Perpinán, M., 2010. Estimating missing data sequences in X-ray microbeam recordings, In: Proc. Interspeech, Makuhari, Japan, pp. 1592-1595.
- Qin, J., Toth, A.R., Black, A.W., Schultz, T., 2008. Is voice transformation a threat to speaker identification?, In: Proc. ICASSP, Las Vegas, Nevada, pp. 4845-4848.
- Repp, B.H., Williams, D.R., 1987. Categorical tendencies in imitating self-produced isolated vowels. *Speech Communication* 6, 1-14.
- Richmond, K., 2001. Estimating articulatory parameters from the acoustic speech signal. Ph.D. dissertation. The Centre for Speech Technology Research, University of Edinburgh.
- Riegelsberger, E., 1997. The acoustic-to-articulatory mapping of voiced and fricated speech. PhD dissertation. Ohio State University.
- Rogers, C.L., Dalby, J.M., 1996. Prediction of foreign-accented speech intelligibility from segmental contrast measures. *The Journal of the Acoustical Society of America* 100, 2725-2725.

- Ryan, E., Carranza, M., Moffie, R., 1977. Reactions toward varying degrees of accentedness in the speech of Spanish-English bilinguals. *Language and Speech* 20, 267-273.
- Ryan, E.B., 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, 855-863.
- Sambur, M., 1975. Selection of acoustic features for speaker identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 23, 176-182.
- Schairer, K.E., 1992. Native speaker reaction to non-native speech. *The Modern Language Journal* 76, 309-319.
- Schroeter, J., 2008. Basic principles of speech synthesis, In: Benesty, J., Sondhi, M., Huang, Y. (Eds.), *Springer handbook of speech processing*. Springer Verlag, Berlin, Germany, pp. 413-428.
- Schroeter, J., Sondhi, M.M., 1994. Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Transactions on Speech and Audio Processing* 2, 133-150.
- Scovel, T., 1988. *A time to speak: A psycholinguistic inquiry into the critical period for human speech*. Newbury House, Cambridge, England.
- Sebastian, R., Ryan, E., Corso, L., 1979. Social judgments of speakers with differing degrees of accentedness, In: *Proc. Ninth World Congress of Sociology, Uppsala, Sweden*, pp. 51-61.
- Sheffert, S.M., Pisoni, D.B., Fellowes, J.M., Remez, R.E., 2002. Learning to recognize talkers from natural, sinewave, and reversed speech samples. *Journal of Experimental Psychology in Human Perception* 28, 1447-1469.
- Shiga, Y., 2009. Pulse density representation of spectrum for statistical speech processing, In: *Proc. Interspeech, Brighton, UK*, pp. 1771-1774.
- Shiga, Y., King, S., 2004. Accurate spectral envelope estimation for articulation-to-speech synthesis, In: *Proc. ISCA Speech Synthesis Workshop, Pittsburgh, USA*, pp. 19-24.
- Simpson, A., 2001. Dynamic consequences of differences in male and female vocal tract dimensions. *The Journal of the Acoustical Society of America* 109, 2153-2164
- Simpson, A.P., 2002. Gender-specific articulatory-acoustic relations in vowel sequences. *Journal of Phonetics* 30, 417-435.
- Sjölander, K., Beskow, J., 2000. Wavesurfer-an open source speech tool, In: *Proc. ICSLP*, pp. 464-467.
- Sjöström, M., Eriksson, E., Zetterholm, E., Sullivan, K., 2006. A switch of dialect as disguise, In: *Proc. Fonetik, Lund, Sweden*, pp. 113-116.

- Stevens, K.N., 1998. *Acoustic Phonetics*. MIT Press, Cambridge, MA.
- Story, B.H., Titze, I.R., Hoffman, E.A., 1996. Vocal tract area functions from magnetic resonance imaging. *The Journal of the Acoustical Society of America* 100, 537-554.
- Strongman, K., Woosley, J., 1967. Stereotyped reactions to regional accents. *British Journal of Social and Clinical Psychology* 6, 164-167.
- Stylianou, Y., 2001. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing* 9, 21-29.
- Stylianou, Y., Cappe, O., Moulines, E., 1998. Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing* 6, 131-142.
- Sundermann, D., Ney, H., Hoge, H., 2003. VTLN-based cross-language voice conversion, In: *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, St. Thomas, Virgin Islands, pp. 676-681.
- Sundström, A., 1998. Automatic prosody modification as a means for foreign language pronunciation training, In: *Proc. Speech technology in language learning*, Marholmen, Sweden, pp. 49-52.
- Syrdal, A.K., Bennett, R.W., Greenspan, S.L., 1995. *Applied Speech Technology*. CRC Press.
- Tang, M., Wang, C., Seneff, S., 2001. Voice transformations: From speech synthesis to mammalian vocalizations, In: *Proc. Eurospeech Aalborg, Denmark*, pp. 357-360.
- Tenenbaum, J.B., Silva, V.d., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319-2323.
- Teutenberg, J., Watson, C., 2006. Vowel quality in accent modification, In: *Proc. Australian International Conference on Speech Science & Technology*, University of Auckland, New Zealand, pp. 292-295.
- Toda, T., Black, A.W., Tokuda, K., 2007. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *Audio, Speech, and Language Processing* 15, 2222-2235.
- Toda, T., Black, A.W., Tokuda, K., 2008. Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Communication* 50, 215-227.
- Tokuda, K., Heiga, Z., Black, A.W., 2002. An HMM-based speech synthesis system applied to English, In: *Proc. Workshop on Speech Synthesis*, Santa Monica, California USA, pp. 227-230.
- Toth, A., Black, A., 2005. Cross-speaker articulatory position data for phonetic feature prediction, In: *Proc. Interspeech*, Lisbon, Portugal, pp. 2973-2976.

- Traunmüller, H., 1994. Conventional, biological and environmental factors in speech communication: A modulation theory. *Phonetica* 51, 170-183.
- Traunmüller, H., 1998. Modulation and demodulation in production, perception, and imitation of speech and bodily gestures, In: *Proc. Fonetik*, Dept. of Linguistics, Stockholm University, Sweden, pp. 40 - 43.
- Turk, O., Arslan, L.M., 2005. Donor selection for voice conversion, In: *Proc. EUSIPCO*, Bogaziçi University pp. 201-204.
- Turk, O., Arslan, L.M., 2006. Robust processing techniques for voice conversion. *Computer Speech & Language* 20, 441-467.
- Vieru-Dimulescu, B., Mareüil, P.B.d., 2005. Contribution of prosody to the perception of a foreign accent: A study based on Spanish/Italian modified speech, In: *Proc. ISCA Workshop on Plasticity in Speech Perception*, London, UK, pp. 66-68.
- Voiers, W., 1964. Perceptual bases of speaker identity. *The Journal of the Acoustical Society of America* 36, 1065-1073
- Watson, C., Kewley-Port, D., 1989. Advances in computer-based speech training: Aids for the profoundly hearing impaired. *Volta-Review* 91, 29-45.
- 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. Report.
- Westbury, J.R., Hashi, M., J. Lindstrom, M., 1998. Differences among speakers in lingual articulation for American English /r/. *Speech Communication* 26, 203-226.
- Wrench, A., 1999. MOCHA-TIMIT. Queen Margaret University College, [www.cstr.ed.ac.uk/artic/mocha.html](http://www.cstr.ed.ac.uk/artic/mocha.html).
- Yan, Q., Vaseghi, S., Rentzos, D., Ho, C.H., 2007. Analysis and synthesis of formant spaces of British, Australian, and American accents. *IEEE Transactions on Audio, Speech, and Language Processing* 15, 676-689.
- Yanguas, L.R., Quatieri, T.F., Goodman, F., 1999. Implications of glottal source for speaker and dialect identification, In: *Proc. ICASSP*, Phoenix, AZ, pp. 813-816.
- Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 1995. *The HTK book*. Cambridge University. Cambridge University Engineering Department 1996.
- Young, S.J., 1993. *The HTK Hidden Markov Model Toolkit: Design and Philosophy*. Department of Engineering, Cambridge University. Report.



## APPENDIX A

### ARPABET

The Arpabet was developed in 1971 by the Advanced Research Projects Agency (ARPA). It represents each phoneme of General American English with ASCII characters. It was selected over the more widely used International Phonetic Alphabet (IPA) because we used HTK and the CMU pronunciation dictionary to perform an initial automatic transcription, which was manually corrected later. The CMU pronunciation dictionary has transcriptions for over 125,000 words. The chart below gives examples for each Arpabet symbol and the corresponding IPA symbol. Arpabet symbols may be listed more than once when a single Arpabet symbol represents more than one IPA symbol.

#### MONOPHTHONGS

IPA	Arpabet	Example	Translation
ɔ	AO	off	AO F
ɑ	AA	father	F AA DH ER
i	IY	bee	B IY
u	UW	you	Y UW
ɛ	EH	red	R EH D
ɪ	IH	big	B IH G
ʊ	UH	should	SH UH D
ʌ	AH	but	B AH T
ə	AH	sofa	S OW F AH
æ	AE	at	AE T

## DIPHTHONGS

IPA	Arpabet	Example	Translation
eɪ	EY	say	S EY
aɪ	AY	my	M AY
oʊ	OW	show	SH OW
aʊ	AW	how	HH AW
ɔɪ	OY	boy	B OY

## R-COLORED VOWELS

IPA	Arpabet	Example	Translation
ɝ	ER	her	HH ER T
ɑ	ER	father	F AA DH ER
ɛr	EH R	air	EH R
ʊr	UH R	cure	K Y UH R
ɔr	AO R	more	M AO R
ɑr	AA R	large	L AA R JH
ɪr	IH R	ear	IY R
aʊr	AW R	flower	F L AW R

## STOPS

IPA	Arpabet	Example	Translation
p	P	pay	P EY
b	B	buy	B AY
t	T	take	T EY K
d	D	day	D EY
k	K	key	K IY
g	G	go	G OW

## AFFRICATES

IPA	Arpabet	Example	Translation
tʃ	CH	chair	CH EH R
dʒ	JH	just	JH AH S T

## FRICATIVES

IPA	Arpabet	Example	Translation
f	F	for	F AO R
v	V	very	V EH R IY
θ	TH	thanks	TH AE NG K S
ð	DH	that	DH AE T
s	S	say	S EY
z	Z	zoo	Z UW
ʃ	SH	show	SH OW
ʒ	ZH	measure	M EH ZH ER
h	HH	house	HH AW S

## NASALS

IPA	Arpabet	Example	Translation
m	M	man	M AE N
n	N	no	N OW
ŋ	NG	sing	S IH NG

## LIQUIDS

IPA	Arpabet	Example	Translation
l	L	late	L EY T
r	R	run	R AH N

## SEMIVOWELS

IPA	Arpabet	Example	Translation
j	Y	yes	Y EH S
w	W	way	W EY

## APPENDIX B

### ARTICULATORY CORPUS TRANSCRIPT

The articulatory corpus transcript consists of 550 unique sentences selected from the Glasgow Herald to provide adequate phonetic coverage. The following 344 utterances were recorded by both our foreign (RGO) and native (MAB) speakers. Immediately following this list are an additional 206 utterances recorded solely by RGO.

1. A breakthrough, I surmised.
2. A few may even have reached Level F.
3. A further eight people are awaiting extradition on terrorism charges.
4. A new school will be built.
5. A new site was also earmarked near the main gate but the plans were never implemented.
6. A popular novel that transcends its genre.
7. Aberdeen is enjoying a buoyant economy, with optimism in the oil and gas industry.
8. Ah, it looks so beautiful, said Noriega.
9. Air fares could rise under proposals to alter the way airports charge airlines for landing and parking.
10. Air to air refueling tankers are also stationed there.
11. Allies reject offer as half measure.
12. America also suffered in the uncertain economic climate.
13. And it is a lie which has done much damage to Catholics.
14. Anxious about commercial oil interests in Iraq.
15. Any political ambitions are also curbed.
16. Anybody with eyes in their head would see that.
17. Are we not worth bothering looking at?
18. As I was driving towards the garage I saw it start to twist and turn.
19. At Edinburgh, my tutors were Robin and Faith Jacques, and Roy Wood.
20. Autism link still cause for concern.
21. Bagpipes rounded off the evening.
22. Baroness Thatcher is the undisputed conference darling.
23. Barrymore flew in to Heathrow airport from Dubai on Tuesday night.
24. Because these deer are gregarious, they go about in groups.
25. Being competitive means coming up with new styles in each genre and constantly trying to be fresh.
26. Big thinking and big money was required.
27. Bill Welsh, chairman, called for a large scale study.
28. Bin Laden returned to Afghanistan.
29. Bodies fall on a nearby golf course.
30. Both failures, two engine failures.
31. Britons warned, foul deed.

32. But everything changes.
33. But he did give himself a severe jolt.
34. But the gurus had got it wrong, and they were wrong now.
35. Change will not occur overnight.
36. Christy Moore covered it on his last album.
37. Club chairman Hugh Scott said yesterday he would defend the action.
38. Cut here, cut there, cut everywhere.
39. Daredevil youth versus experience.
40. Each case is unique.
41. Elsewhere, change is mixed.
42. Eriksen saw out the rest of the war as a detainee and was then released.
43. Even on Air India, our fleet is very old.
44. Everybody meddles with nature.
45. Extra officers arrived late last night.
46. Fears of sabotage were rife last night.
47. For good measure he offered an unreserved apology.
48. For now our trip as we knew it is over.
49. Four year beef would cost too much.
50. Good to know, Joe.
51. Gordon Brown was superb.
52. He added: She sure knows her rugby.
53. He adored Whoopi and the two were great friends.
54. He claimed that racism drove him out of Scotland.
55. He did not make it to L A.
56. He has been sighted inside the French zone, but no attempt has been made to apprehend him.
57. He is action oriented.
58. He is just a stopgap governor, said a spokeswoman for the Stop The Closure campaign.
59. He is well placed to make such judgements.
60. He later met Jacques Chirac, the French president.
61. He now moves up to fifth place.
62. He said that the only figures that were down were for international leisure visitors.
63. He said: Education education, education.
64. He said: Nigel went to L A in January and acted in a film with Whoopi Goldberg called Call Me Claus.
65. He saw it all as a big game.
66. He was attached to the Texas militia air force.
67. He was flirting, touching my arms.
68. He was sacked last year amid nepotism allegations.
69. He wore dark jeans and a sweatshirt.
70. His Del Boy pitch was well rehearsed.
71. His job was not advertised.
72. His number two has been confirmed as Dick Cheney, another Texan oil millionaire.
73. His strange, smooth jowls wobble like those of a barnyard turkey.
74. Hispanic costumes are quite colorful.
75. Hopefully we are catching up.
76. Hospital routine would go out the window.
77. How long does it go on? They have not made their target so far.

78. How wrong we all were.
79. However, friends said he had come north only two weeks ago.
80. However, the report did give some idea, albeit warning against comparison.
81. However, the youngster died of exposure shortly before the party was rescued.
82. However, there is now artistic concern.
83. I also go up north to hunt deer, he said.
84. I always believed it myself.
85. I am gullible, he replied.
86. I am not a witch, she said.
87. I am now unemployed.
88. I am sure they all work very hard.
89. I chose coronation chicken.
90. I enjoy going to work.
91. I feel genuinely lucky.
92. I grew up in L A, and I know the level of the beating down that happens to film makers.
93. I have enough vices without smoking dope.
94. I have warm Portuguese blood.
95. I just had to be here, she said.
96. I know what sectarianism can do.
97. I look forward to others following suit.
98. I never saw anything.
99. I never thought in a million years about terrorism, she says.
100. I often produce tears.
101. I opened the shutters and saw a man.
102. I reject this accusation.
103. I think any such policy should also specify how it would use any actuarial information.
104. I urge you to opt for restraint.
105. I was directed to destroy all copies and wipe the computer file.
106. I worked in retail for years and enjoy shopping.
107. I would also urge people to avoid stocking up.
108. If children behaved like this they would get a smack round the ear.
109. If the E U does not heed W T O rulings, there will be mayhem.
110. If they insist on breaking the law, I would hope that the courts would hit them as hard as they can.
111. If you want to regulate noise, regulate noise.
112. In Paris, Jacques Chirac, the president, told his armed forces to be prepared for deployment.
113. In the dry season, each village seems a pretty idyll.
114. Iraq had, for example, a programme to modify aerial fuel tanks for Mirage jets.
115. It can be used against both bribe givers and takers.
116. It can change from hour to hour.
117. It does look very eerie and that is one of the reasons I love it.
118. It features the job title Generator.
119. It has been a fascinating job, he said.
120. It has provided employment in a lot of rural areas where there was no employment, he said.
121. It involved nine thousand pounds for Stirling Royal.
122. It is about the prism through which we view the world.

123. It is due for release in the U K early next year.
124. It is not hers.
125. It is time to change our approach.
126. It is, of course, the English theme pub.
127. It manufactures hair products.
128. It seems wrong that a tiny little office in Edinburgh exerts this control over the output of Scotland.
129. It was a great job but times have changed.
130. It was a magazine of leisure, not war.
131. It was carnage, carnage.
132. It was like waving a red flag, she said.
133. It will be a superb gateway attraction to the Highlands.
134. It will set tourism back a year.
135. Jacques Chirac has our support.
136. Last month oil prices were at their highest since the Iraqi invasion of Kuwait in nineteen ninety.
137. Last night they unearthed a seven year old boy and his mother.
138. Lauren has a twin sister, Lisa Ann.
139. Leisure centres were also affected.
140. Leisure: thirty three percent expect Mexico and Australia to become regular holiday destinations.
141. Living organisms and parts of living organisms are not inventions.
142. Louise Hopkins, visual arts.
143. Make sure you know where the exam is being held.
144. Many are now orphans.
145. Meanwhile, the only birdies on the luxuriant new course are gulls from the Firth of Forth.
146. Michael Ashcroft is a British citizen.
147. Michelle saw Arlene the next day.
148. Military courts always apply English criminal law, even for trials in Scotland.
149. Morgue vans are more common and their numbers are increasing.
150. Most of his works are painted using oils or watercolors.
151. Mr Hague resubmitted the nomination earlier this year.
152. Mr Hague, trailing so badly in the polls, certainly put on a bravura performance under the circumstances.
153. Mr Howie said: Burns is alive and kicking.
154. Mr Mustafa had to help bury them.
155. Mr Shevardnadze was enraged.
156. Mr Straw urged all politicians to moderate their language and declined to condemn Mr Hague.
157. Ms Doherty, who was on her first protest, is the only British female being held.
158. Ms Jackson can vouch for that.
159. My brother in law only has me, but what about the people who do not have anybody?
160. Nato authorizes air strikes.
161. No one has yet come up with answers.
162. None more important to that soil.
163. Nor did he take time off for leisure pursuits.
164. Not the overcoat surely, I thought.



165. Nothing more, nothing less.
166. Now I have something else to aim for.
167. Now we always have fun here.
168. Now, get out of here.
169. Oasis left on a real high.
170. One convoy had four buses.
171. One person jumped into the water.
172. Oops There goes another one!
173. Or are the metaphor mongers getting carried away with themselves?
174. Oswald Road is situated in the Grange, a conservation area.
175. Others are obtained from nursing agencies, which charge a commission for placing nurses.
176. Our book shows you how.
177. Our budget could not bear that.
178. Our percentage vote is higher than the Conservative percentage vote.
179. Our proposals tackle areas of real need.
180. Our reasons are not entirely driven by the low oil price, although the volatility is a factor, he said.
181. Our research points the way.
182. Part of England, is not?
183. People ask me how I can do it.
184. People buy everything with cash.
185. Politically, the system is a hot potato, particularly following devolution.
186. Porcupines resemble sea urchins.
187. Recovery was soon aided by a bottle of diet lemonade.
188. Robb, thirty one, was granted a prison transfer home to Northern Ireland in nineteen ninety seven.
189. Sadly, they are usually all of these things.
190. Seventeen eighty five: John Jeffries and Jean Pierre Blanchard cross the Channel in a balloon.
191. Shaving cream is a popular item on Halloween.
192. She always jokes about too much garlic in his food.
193. She became pregnant with their son Ross, now aged five.
194. She died in middle age.
195. She has always been a healthy, strong person.
196. She said of the vandalism: I am shocked, and a bit scared.
197. She said: Seizure rates are the tip of the iceberg.
198. She survived because she had been able to shelter in a bivouac bag.
199. She took days off work.
200. She was vague, just generally vague.
201. She went into hospital early last month after cancer reappeared.
202. Sir John Kerr is a man who oozes diplomacy from every pore.
203. Six Scottish plants employ thirteen thousand.
204. So I urge the world to finish the job.
205. So now you know.
206. So too is the other Carry On Doctor alter ego, the Hattie Jacques battle axe.
207. Some had their eyes gouged out or heads smashed in.
208. Sometimes in order to save the organism you have to sacrifice a limb, she said.

209. Soon after, it clouded over with dark clouds.
210. Special measures were needed.
211. Still more stark is the threat of more terrorism on American soil.
212. Such a move might torpedo any peace moves.
213. Take the issue of poverty.
214. Thank you, King Neptune.
215. That idea is rubbish, he said.
216. That may yet prove to be one closure too far.
217. That noise problem grows more annoying each day.
218. That party, United Torah Judaism, also broke ranks in the budget vote.
219. That was their final wish, she said.
220. The British government should ensure by what they say and what they do.
221. The Cass report outlines five options for the company if closure goes ahead.
222. The Chinook was plagued with problems.
223. The Ethiopian population is rising by around three percent per year.
224. The Gucci studio recently moved there.
225. The Gujarat state government had a more conservative estimate.
226. The animal also died.
227. The arthritic knees gave way.
228. The boy vanished as his mother was putting his sister in her car seat at the visitor centre.
229. The butcher is very good.
230. The chancellor has certainly had to backtrack on the main forecasts he made in his April budget.
231. The city is thriving economically.
232. The double whammy for Da Ali G Show was also significant.
233. The end of the case creates a kind of closure for the family.
234. The executive version was finally carried.
235. The film was a sensation, and may have helped change the law on homosexuality.
236. The fourth area is public buildings.
237. The main body of troops is not due to deploy until next Friday at the earliest.
238. The man called the garage earlier this week after seeing the car advertised on the web.
239. The oasis was a mirage.
240. The obvious answer is cash.
241. The officer said he went to see the boy after that time, and could see no marks on his face.
242. The patch up jobs have come and gone.
243. The person who wrote that should be horse whipped.
244. The pressures are intense.
245. The prime minister dismissed any suggestion that the war was in any way to do with oil.
246. The proper guests were offered a finger buffet.
247. The second sort of love is the Mentor and the Protege, or Pygmalion complex.
248. The smoke, she says, swept through the narrow canyons like a tsunami.
249. The teams will use radar equipment employed by oil companies to track underground cables.
250. The trademark cowboy boots are the giveaway.
251. The unions are powerless except in nooks of the public sector.
252. The vibe, the buzz in Glasgow, is amazing.
253. The war on terrorism was going global.

254. The woman sits down, fuming.
255. The words Algeria and terrorism sit too easily together.
256. Then he made it into Big Oil.
257. There are garages below me.
258. There are mechanisms to shine a light.
259. There are seven other contenders.
260. There is confusion when more of the Opposition arrive.
261. There is no clear idea on where all this fresh water is going to come from.
262. There is no logical reason for that.
263. There is no switch off.
264. There was huge irony here.
265. There was no answer.
266. There were no casualties.
267. These are fairly rare.
268. They are all dodging the issue.
269. They are so easy for youngsters to open.
270. They are standing idle.
271. They are thoroughly evil people.
272. They entered everywhere.
273. They feel let down by them.
274. They know the truth, they really know the truth.
275. They lie at the strategic heart of the new oil bonanza.
276. They now have a clear choice, he said.
277. They now have jobs with U K Coal.
278. They play a key role in shaping young lives.
279. They remained lifelong friends and companions.
280. They wanted to know why the names Utah, Omaha, and Overlord had appeared in his answers.
281. They would have been booked very early.
282. This is a virtuoso Chancellor with unprecedented intellectual command at the height of his powers.
283. This is the big fear.
284. This may have been a deliberate ploy on the part of Colonel Cody.
285. This should allow aircraft to move more smoothly across European air space, he said.
286. This was a meeting which changed his life.
287. This was always a false choice for New Labour, he argued.
288. Those chefs know who they are.
289. Three others were wounded.
290. Time is now very short.
291. To fearsome and feisty now add fast.
292. Tourism has changed all that.
293. Two Asian youths were seen running from the scene.
294. Two bows were also stolen.
295. Vandalism also leapt thirteen point three percent.
296. Virtually all genres of musical taste are recognized in the poll.
297. We are considering the options.
298. We are curious to see if a Scottish genre emerges.
299. We are not dealing here with refugees, he said.

300. We are not working with theories.
301. We are now dealing with Bin Laden.
302. We are nowhere near there yet.
303. We are open every Monday evening.
304. We are regarded as being dour people.
305. We can enjoy alcohol in other areas of life but not in church premises.
306. We depend on tourism for our livelihood.
307. We desperately need the concierge back.
308. We enjoy every minute we have with her.
309. We have to ensure victims are protected.
310. We just feel fear, fear for her safety.
311. We must all play our part.
312. We must take a measured look at this.
313. We need to reach out to all areas of our society.
314. We only learned at lunchtime when she phoned home that she actually was.
315. We shall take very serious measures.
316. We will also be using the final to highlight our tourism potential, he said.
317. We will do our best to ensure that the tourism areas are not too badly affected.
318. We will drop five thousand tons a month into drop zones.
319. We will now pursue this objective with renewed vigor.
320. We will raise the minimum income guarantee in line with earnings next year.
321. We would be happy if there was zero compensation.
322. We would rather have guidance now.
323. We would really like to see more money being earmarked to prevent any possibility.
324. West coast and Cross Country drivers now earn thirty two thousand five hundred pounds.
325. What he put people through.
326. When all else fails, use force.
327. When he saw us, he ran back inside and we followed.
328. When is a resolution not a resolution?
329. When these two K G B returned to Moscow, they were tried on espionage charges and executed.
330. Where things can be saved, that will be done.
331. Which is to secure shipbuilding on the Clyde.
332. Who are the bunglers now?
333. Whose side is the law on: those with a position in life or the victims?
334. Why on earth should we do that?
335. Will you renew this worthless vow in this election?
336. Winners receive a redcoat statuette.
337. With Lee, I had high blood pressure.
338. Yields will be highly variable.
339. You are from Scotland?
340. You enjoy your flight?
341. You have not done your job properly.
342. Yours sincerely, Elizabeth R.
343. Zimbabwe is the first example.
344. Zoo authorities are considering whether to put her down.

RGO additionally recorded the following 206 utterances. All 550 sentences took 6 hours to record and is the most extensive single session articulatory corpus to our knowledge.

1. A Labour nightmare.
2. A crusade against a real social evil Leader.
3. About your boy?
4. Absolutely perfect.
5. Age thirty eight.
6. All of sudden, it started opposing terrorism under pressure.
7. All the world has now seen the footage of an Iraqi Mirage aircraft with a fuel tank modified to spray biological agents over wide areas.
8. All wore beards.
9. Among those joining Crook, Hughes, and Bailey in the Scottish version of Guttled! will be Dylan Moran, the youngest Perrier award winner at the Fringe, Daniel Kitson, a Perrier winner last year, and Michael Redmond, of Father Ted fame.
10. An additional four million pounds was withdrawn from D A Afghanistan Bank in Kabul a month after the U S airstrikes began, and none of it has been seen since.
11. An estimated fifty thousand Scottish and Irish sea birds died in the Prestige oil disaster off the Spanish coast last November, it was revealed yesterday.
12. An interpreter was provided and evidence was taken in Hindustani.
13. And the Bradbury Building in L A, an office building, has a fantastic gothic interior that has been seen in Blade Runner, Wolf, Caprice, Murder in the First Degree and loads more.
14. Appeals continuing, but confusion over the secretive judicial process.
15. Are harmful.
16. Are you all right?
17. Are you pregnant?
18. Arriving early, I looked around Bournemouth.
19. Auctioneers: United Auctions.
20. Audio tape played.
21. B A asked all crews on flights in U K airspace to invite passengers to observe the silence in the air.
22. Bachelor of Arts with Honors: Clare Ritchie.
23. Backie was dead.
24. Barcelona eight.
25. Battle for power.
26. Because in Scotland we are offering education in art and design to little more than half the proportion of students than in England, and even the English are not particularly generous in providing for art education.
27. Best German Act: Guano Apes.
28. Biography Churchill, by Roy Jenkins.
29. Both China and India.
30. Both claim it is too early to tally up.
31. But surfers beware.
32. Carlisle.
33. Change.
34. Charles Macintosh: Raincoat.

35. Clear my name.
36. Club fears.
37. Collision alarm off July forth.
38. Countdown to conflict, February forth.
39. Cutting edge Leader.
40. Cynicism is corrosive.
41. David D.
42. Death Breath, Thunder Thighs, Jug Ears, Baldy.
43. Diana.
44. Die, pig, die.
45. Dirt will fly.
46. Distinction: Cameron R Yates.
47. Dressed in turquoise with a matching hat, she delighted onlookers at the kirk as she arrived in a black limousine.
48. Education lesson.
49. Eh?
50. Eight: Belinda Earl, forty, chief executive of Debenhams.
51. Ethel Elizabeth Nunn, Founder, Society of Friends of the Lotus Children.
52. Even Yom Kippur was a blow because of initial enemy advances in the Sinai desert and on the Golan Heights.
53. Failure to deliver Death knell.
54. Fire tenders moved into position as the Concorde landed at Gander, where B A engineers are still inspecting the plane.
55. Fish like the salmon in the River Loire will disappear, while desertification of already arid areas, like central Spain, will increase.
56. Five years later, he re entered the world of business with the creation of his own engineering company aided by the gift of a start up fund from Ms Bourgeois.
57. For services to Lothian and Edinburgh Enterprise.
58. Forget the gormless goldfish jokes.
59. Forget the row over G M food.
60. Forgotten victims.
61. French President Jacques Chirac said the air attacks were launched to defend peace on our soil, peace in Europe.
62. George W.
63. Germany reported five more deaths yesterday, putting the toll there at seven, and raging waters cut off some towns in the state of Saxony.
64. Given that many newspapers bring oodles of staff, that means lots and lots of fifteen pound donations into Labour coffers.
65. Glasgow.
66. Hague the nationalist.
67. Happy Birthday merger.
68. Has let Mr Mugabe too often off the hook.
69. Having a background in both Urdu and English has probably helped because I understand language structure better.
70. He read zoology at Dundee University and then in Edinburgh, but dropped out of both courses.
71. Health Which? offers a range of actions individuals can take to reduce risks.
72. Heseltine turns on Hague; Heseltine in focus.

73. Him very often at my check out.
74. Hip, hip, hooray!
75. Hope and skepticism editorial comment.
76. Horrifying fetish, bizarre behavior.
77. Hotel California Eagles.
78. How can we ignore it?
79. How far do you go? He said he would not be calling for other customers to follow their lead, but would leave them to form their own opinion.
80. How serious it is.
81. However, she was passed over for the Chinese team, had given up the sport, and was at Wuhan university when she came to Britain to study English.
82. However, the demand for heavy, hi tech armored leviathans is dwindling.
83. However, there was one faux pas on greeting a well known businessman, accompanied as usual by a female on his arm.
84. However, this is set to decline as pensions become less generous, with the report noting: The current batch of retirees may have some of the highest proportions of those who can enjoy early retirement with a relatively high level of income.
85. I A, U S astronaut with Scots ancestors, yesterday predicted the future of space exploration could include unearthing the mysteries of Mars, a return to the moon and the development of the Hubble space telescope.
86. I am getting air, but ...
87. I appreciate that.
88. I believe ruthenium compounds could one day be suitable for clinical use, and have a significant impact on drug development against cancer.
89. I enjoy reading.
90. I like girls.
91. I.
92. In addition, there were all those passengers wanting to come ashore and enjoy Ayrshire.
93. In his post, Mr Jung will oversee the teaching of environmental art, painting, photography, printing and sculpture.
94. Iraq must disarm.
95. Is she alive?
96. Is she dead?
97. Is she in trouble?
98. It is half baked.
99. It jars.
100. It never occurred to me that a rule would be necessary to keep racism out of blood transfusions and donations.
101. Jane Eyre eleven.
102. Japanese oysters have also started to breed in British waters, threatening to push out the U K oyster in the future, the report said.
103. July eighteenth.
104. July forth.
105. June forth.
106. Jurassic Park nineteen ninety three: eighteen.
107. Lanyard wars, part three.
108. Last month Yves Saint Laurent said in a veiled attack on McQueen that there are murderers in the couture houses.

109. Leader May eighteenth.
110. Leaders from the majority of European countries also made a point of expressing their support, including Jacques Chirac, the French president.
111. Legal action April sixteenth.
112. Mafia gangs.
113. March first.
114. March ninth.
115. March seventh.
116. March sixth.
117. March third.
118. March.
119. Missiles are jeopardising aid operations.
120. Moby Grape Moby Grape.
121. Model chamber.
122. Moon Safari Air.
123. Ms Chen has developed special organic coats for ruthenium one of the rarest metals on earth which enable compounds to target D N A bases of cancer cells with greater accuracy.
124. Must change channel.
125. My idea of the winner?
126. Naomi Mitchison loved people.
127. Neighbors falling out.
128. New era delays.
129. Nick Nairn China Surprise!
130. No known cures.
131. No sunshine break.
132. Normal.
133. Nothing, she replied.
134. O: Have you located the wound yet?
135. October seventeenth, two thousand Hatfield.
136. Oh dear.
137. Oh, yes, I vote.
138. Oil embargo endorsed.
139. Or depths.
140. P R?
141. Parliamentary sketch: Murray Tosh profile.
142. Partial female triumph.
143. Poor chap.
144. Prison Aberdeen.
145. Prison Garth.
146. Push off.
147. Reserve League.
148. Rod Eddington, B A chief executive, recorded a message for staff explaining that action was needed to lead B A out of crisis and return the airline to profitability.
149. Samuel Gamble, School Caretaker.
150. Sarah Lancashire seven.
151. Sharp.



152. Smashed tumblers, dropped forks and, the bane of Scottish goal keepers, a goalmouth fumble.
153. So I chose Belgium.
154. Sort of.
155. Stay off her.
156. Tackling Al Qaeda is the equivalent of fighting shadows.
157. Take Northern Ireland.
158. The Ayrshire line was reopened during the early evening rush hour.
159. The Big Yin is celebrating the arrival of the Wee Yin.
160. The Pentagon has drawn up a blueprint for a shock and awe air assault on Iraq which will concentrate on killing as many of its leaders as possible and cutting the survivors off from contact with their troops in the field.
161. The Tweets are worthy winners.
162. The choice is clear.
163. The empty vessel was due to collect plutonium fuel from the La Hague nuclear reprocessing plant near Cherbourg for shipment to Japan.
164. The farther north you go, the more bookish folk become, with Orkney tops in Britain with fourteen withdrawals per person per year.
165. The global villain, George Bush, has been foiled again.
166. The gorgeous butterfly ate a lot of nectar.
167. The harsh assessment from the European Commission came hard on the heels of criticism earlier this week of Scottish beaches in a separate British report.
168. The idea I got was that the U K had very good people, very hospitable people, but now my idea regarding the people has completely changed.
169. The idea I got was that the United Kingdom had very good people, very hospitable people, but now my idea regarding the people has completely changed.
170. The official said Indian troops had returned the firing and the exchange had continued in the Uri sector of the control line.
171. The political battleground over police numbers and crime figures escalated fiercely in Westminster amid bitter exchanges among politicians.
172. The star made a series of disclosures in the one hour, fifty minute programme about his life, his appearance and his three children his five year old son, Prince Michael the first, daughter Paris, four, and his baby boy.
173. The two leaders are due to meet close to where Mr Blair and his family are staying near Toulouse, in southern France.
174. The vital Alaska oil pipeline, which carries crude oil from northern Alaska to terminals at Prudhoe Bay, is being monitored by helicopter patrols, while the National Guard is protecting the plant in Michigan that makes anthrax vaccine for the military.
175. The water was slowly oozing out a soft and transparent light it must have accumulated the day before or, more likely, during the last summer.
176. There have been shows in Japan, the U S, and a major art nouveau exhibition at the Victoria and Albert Museum in London.
177. There is a wider public concern about the gradual shift from N H S care, which is free, to social care, which is means tested.
178. There was no charge.
179. They are going after bin Laden.
180. They love modern art.
181. This gave some hope.

182. This will make job creation even more difficult, Mr Duncan Smith said.
183. Thus far.
184. To ensure genuine football fans can enjoy their day without fear for their safety.
185. Treasure Island Robert Louis Stevenson: two.
186. Tumultuous cheers.
187. Typical old style Scottish Labour M P.
188. U K French summit, where Tony Blair will try to win round a previously skeptical Jacques Chirac, the French president.
189. Ugh!
190. Until last summer, the force had turned a blind eye to prostitutes using the Leith zone for almost two decades.
191. Voila!
192. Watch this space.
193. Watch your back!
194. We are Chechens.
195. We are frustrated because we are now in the third year of a major recruitment initiative to encourage more applicants from the state sector and this kind of story can do nothing but sabotage our efforts, she concludes.
196. Weird kind of beauty.
197. Well, more fool me.
198. What Tony Blair thinks.
199. What emerges is a world in which every idea, every device is harnessed to meet company needs: more smoking by more people.
200. What is she doing?
201. Why do I say this?
202. Why should she be?
203. Wrap up warmly.
204. Wrong.
205. You know what I feel?
206. You over in States on business, sir?

## APPENDIX C

### ISOMAP

To perform multidimensional scaling, we first create a (100×100) matrix containing the average perceptual distance between any two of the 100 utterances. Shown in Figure 43 (a) as an image (darker colors indicate larger perceptual distances between the corresponding pair of utterances), this matrix is sparse due to the large number of utterance pairs (10,000) relative to the number of participants. To guard against outliers, we eliminate any utterance pairs that have been rated by only one participant. We use an  $\varepsilon$ -neighborhood with a radius of 7 perceptual units<sup>25</sup> to define a local connectivity graph; the resulting local distance matrix is shown in Figure 43 (b). Geodesic distances between every pair of utterances are then estimated using Dijkstra’s shortest paths algorithm (Dijkstra, 1959), which results in the fully connected distance matrix  $D$  shown in Figure 43 (c).

Following Tenenbaum et al. (2000), we apply an operator  $\tau(\cdot)$  to matrix  $D$ , which converts distances into inner products:

$$\tau(D) = -\frac{HSH}{2}$$

where  $S$  is a matrix containing the squared distances found in  $D$  (i.e.  $S_{ij} = D_{ij}^2$ ),  $H$  is the centering matrix

$$H = I_N - \frac{1}{N}$$

---

<sup>25</sup> Scores from zero to seven indicate pairs of utterances that participants believed to have been produced by the same speaker.

$I_N$  is an identity matrix, and  $N$  is =100. The  $i$ -th component  $y_i$  of the  $d$ -dimensional embedding (i.e., the coordinates of the  $N$  utterances on the  $i$ -th dimension of the embedding) is found by

$$y_i = \sqrt{\lambda_p} v_p^i$$

where  $\lambda_p$  is the  $p$ -th eigenvalue of the matrix  $\tau(D)$  and  $v_p^i$  is the  $i$ -th component of the  $p$ -th eigenvector. Each of the 100 samples is then represented in two dimensions as  $(y_1, y_2)$ . A 2-dimensional embedding of the distance matrix in Figure 43(c) is shown in Figure 21 (b).

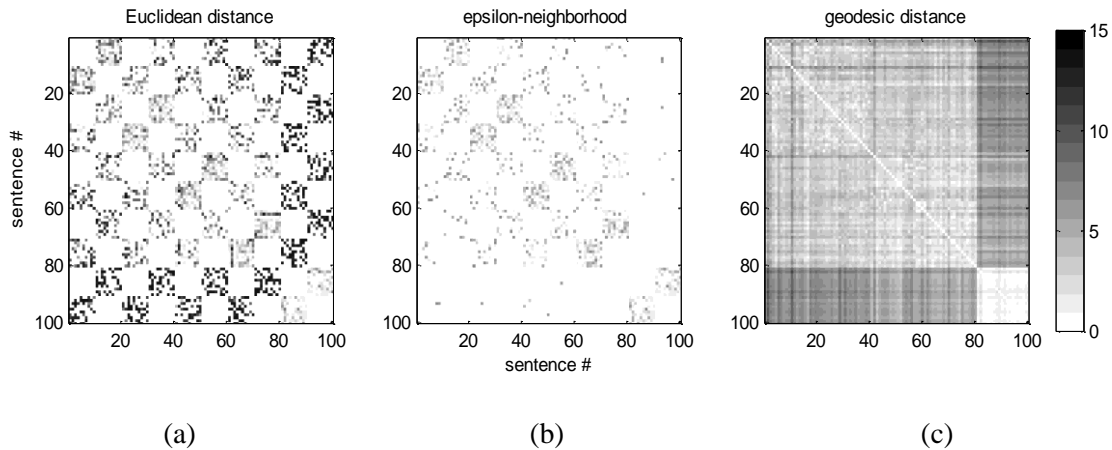


Figure 43 Calculating a complete set of geodesic distances from a subset of utterance pairings. Utterances are displayed in groups of 20, corresponding to their stimulus condition (i.e., utterances 1-20 are from condition 1, 21-40 from condition 2, etc.) Dark pixels indicate that (on average) the corresponding pair of utterances was perceived as having been produced by different speakers; the grayscale is shown on the far right. (a) Raw average distances for the identity experiment with reversed speech. A checkerboard pattern appears due to the testing procedure<sup>26</sup>. (b) Data is thresholded to remove distances greater than seven; this separates pairs of utterances that were perceived as “from the same speaker” from those perceived as “from different speakers”. Utterance pairs for which data was scarce (less than two examples) were also removed to avoid potential problems with outliers. (c) Fully connected graph reconstructed by Dijkstra’s shortest path algorithm. Notice the block structure showing low geodesic distance within utterances from condition 5 (native) and within utterances from conditions 1 through 4 (non-native glottal excitation). It is this distribution of geodesic distances that leads to the clusters observed in Figure 21 (b).

<sup>26</sup> The 20 distinct sentences were divided into two sets (1-10 and 11-20) to ensure that pairs were linguistically unique. Presentation was counterbalanced across sets (i.e. a sentence from the first set was not always played first).

## APPENDIX D

### PROCESSING ARTICULATORY DATA

Electromagnetic articulography (EMA) provides a rare view of the human vocal tract in motion, but it is not perfect technology. Two of the most critical limitations are long term drift of the sensor values and missing data. This appendix describes techniques to address these issues. The articulatory corpus used in this study was collected in a similar manner to the one used in Korin Richmond's dissertation on articulatory inversion (2001). He performed a long term analysis of the pellet positions by calculating their mean value for each utterance in the database. We expect these values to be randomly distributed around the true mean due to the distinct phonetic content of each utterance, however, Richmond's analysis revealed that pellet locations gradually drifted; this effect is also present in the RGO and MAB datasets. As prescribed by Richmond, we estimate a drift offset for each sensor by low-pass filtering (with normalized<sup>27</sup> cutoff at 0.04) the average utterance values (Figure 44).

---

<sup>27</sup> The cutoff for a normalized filter is expressed as a value between [0,1], with 1 corresponding to the Nyquist frequency of the signal. We cannot express the cutoff in absolute terms (e.g. Hz) due to the fact that each sample corresponds to an average value across an utterance and these are not equally spaced. However, if we assume the average duration of utterances to be 2 seconds, then a normalized cutoff of 0.04 corresponds to an absolute cutoff frequency of 0.01 Hz.

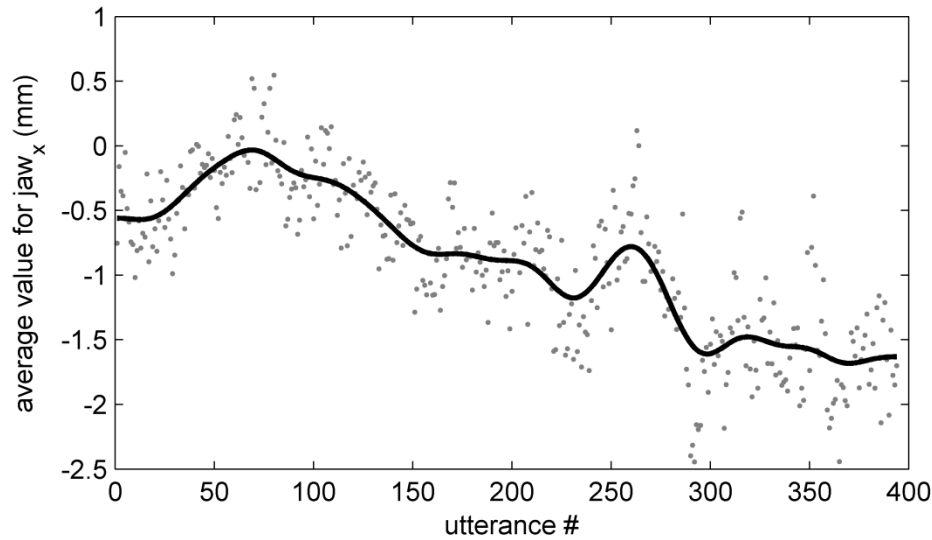


Figure 44 Long term drift correction. Each gray point represents the average value per utterance of MAB's horizontal jaw position. Local variation can be attributed to varying phonetic content, but long-term variation suggests sensor drift.

Occasionally EMA will lose track of a transmission coil for a few hundred milliseconds. Our corpus contains at least one missing pellet in 2% of the samples, but it is rare to find a sample that loses track of all sensor locations. Rather than discard the entire sample, we reconstruct the missing values by exploiting the highly correlated nature of the data. In general, this process is known as data imputation (Little and Rubin, 1987). We employ a method developed specifically for articulatory datasets, which has the ability to represent multi-modal data (Qin and Carreira-Perpinán, 2010). The approach first estimates the joint distribution for all complete samples in the corpus (i.e. no missing values) using a Gaussian mixture model. Missing values are then filled with the maximum likelihood estimate of the conditional distribution specified by the known values. The method has been shown to reconstruct artificially missing data within 1.5 mm (Qin and Carreira-Perpinán, 2010).

## **APPENDIX E**

### **MECHANICAL TURK TESTS**

This appendix provides samples of the web-based evaluation forms used in the ConFAC experiments (Section 7). These forms were hosted by Amazon's online crowdsourcing tool: Mechanical Turk. Participants were paid \$1 for their involvement. All participants were required to pass the American Accent classification task (Figure 45) before being permitted to take part in a real experiment. This was performed to increase the probability that our subjects were native speakers of American English. The standard accent rating task used in Experiments #1 and #2 is shown in Figure 46. The third ConFAC experiment was a forced choice experiment which required participants to select the more natural and intelligible utterance (Figure 47).

### American accent test

The first portion of this qualification gathers personal information.

---

Is American English your native language?

yes

no

---

Do you have any speech impairments?

yes

no

---


Do you have any hearing impairments?

yes

no

---

1. Which regional accent matches the speaker below?




General American

Southern

Northeast

---

2. Which regional accent matches the speaker below?




General American

Southern

Northeast

---

3. Which regional accent matches the speaker below?



General American

Southern

Northeast

Figure 45 The pre-qualification test. Potential participants were required to correctly identify eight out of ten regional American accents before they were allowed to serve as a subject on a real experiment. This initial screening was performed to increase the probability that our subjects were native speakers of American English.



## Accent rating task

This HIT involves classifying the degree of foreign accent for each of the sentences below. For example, Arnold Schwarzenegger is "extremely" accented while Barack Obama is "not at all" accented.

If you are not a native speaker of American English, please include your native language/dialect in the comments box at the bottom of this page - it will NOT affect your reward.












Listen	Not at all accented		Somewhat accented		Quite a bit accented		Extremely accented
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 46 The accent rating form used in ConFAC experiments #1 and #2 (sub-sections 7.2-7.4). Clicking the listen object played a single utterance. Subjects were asked to rate each utterance on the degree of accentedness.

## Evaluation of a speech synthesizer

This HIT is part of a research project investigating the perception of synthesized speech. Your task is to select the best example of a given sentence. Please consider the quality, naturalness, and intelligibility of sentences as a whole. This test should take about 10 minutes.

In each example, after clicking the play button, you will hear two synthesized sentences. Most questions will contain 2 examples of the same sentence, but a few questions will contain different sentences like this one:

Target sentence	Listen	Prefer 1	Prefer 2
Hospital routine would go out the window.	<input type="button" value="▶"/>	<input type="radio"/>	<input checked="" type="radio"/>

This type of question only has one correct answer (i.e. the second choice) and is used for quality control. You must answer the majority of these questions correctly to be paid. Thank you for participating in this study.

Target sentence	Listen	Prefer 1	Prefer 2
It can be used against both bribe givers and takers.	<input type="button" value="▶"/>	<input type="radio"/>	<input type="radio"/>
For good measure he offered an un reserved apology.	<input type="button" value="▶"/>	<input type="radio"/>	<input type="radio"/>
Mr Mustafa had to help bury them.	<input type="button" value="▶"/>	<input type="radio"/>	<input type="radio"/>
There is no logical reason for that.	<input type="button" value="▶"/>	<input type="radio"/>	<input type="radio"/>
I just had to be here, she said.	<input type="button" value="▶"/>	<input type="radio"/>	<input type="radio"/>
They are so easy for youngsters to open.	<input type="button" value="▶"/>	<input type="radio"/>	<input type="radio"/>
At Edinburgh, my tutors were Robin and Faith Jacques, and Roy Wood.	<input type="button" value="▶"/>	<input type="radio"/>	<input type="radio"/>
This was always a false choice for New Labour, he argued.	<input type="button" value="▶"/>	<input type="radio"/>	<input type="radio"/>
Those chefs know who they are.	<input type="button" value="▶"/>	<input type="radio"/>	<input type="radio"/>
Everybody meddles with nature.	<input type="button" value="▶"/>	<input type="radio"/>	<input type="radio"/>

Figure 47 The evaluation form for the third ConFAC experiment. Clicking a single listen object plays two consecutive utterances, allowing us to control the order of presentation. Subjects were asked to select the more natural and intelligible utterance.

## **APPENDIX F**

### **CONFAC TOOLBOX**

The Concatenative Foreign Accent Conversion (ConFAC) toolbox is a speech synthesis tool developed at the Texas A&M University PRISM laboratory. The main functionality supported by this tool is to perform unit selection speech synthesis. The software also comes with two “voices”: an American English speaker and a Spanish accented speaker. ConFAC can perform crude text-to-speech synthesis, but it was designed to alter characteristics of speech related to accent. This appendix contains an abridged version of the ConFAC user’s manual.

#### **GETTING STARTED**

ConFAC was developed on a quad-core machine (i5, 4 @ 2.67 GHz) with 4 GB of RAM running (64-bit) Windows 7. It was coded entirely in the Matlab programming language (2010b) with the following Matlab toolboxes: Control, Image Processing, Neural Network, Optimization, Signal Processing, and Statistics. Several third-party toolboxes are also provided on the installation DVD.

#### **SOFTWARE INSTALLATION**

The included DVD contains additional third-party toolboxes and custom code developed specifically for ConFAC. Installation consists of copying the entire contents of the DVD to a computer and adding all directories and subdirectories to the Matlab path. For example, copy the contents of the DVD to C:/ConFAC. Startup Matlab and run the commands:

```
addpath(genpath('C:/ConFAC'));  
savepath;
```

The “savepath” command saves the new path for future Matlab sessions.

### *Corpus installation*

Before ConFAC can be used, the speech corpora need to be installed. Copy the RGO and MAB datasets from the installation DVD to your computer. ConFAC is a unit selection synthesizer that requires accurate phoneme labeling to produce a good result. Initial phonetic alignment was performed automatically using HTK and then refined manually using Wavesurfer (Sjölander and Beskow, 2000). Phone labels are specified in a standard HTK \*.lab format in each corpus. Mispronunciations in RGO are specified using a third \*.lab file as follows. For every parallel utterance that was recorded for RGO and MAB, there exists an additional \*.lab file that aligns MAB's pronunciation to RGO's acoustic waveform. Mispronunciations can be easily detected since this lab and the true RGO lab have a common frame of reference; the mispronunciation files are stored in a "mab" subfolder inside the main RGO corpus location.

Raw EMA recordings must be processed to estimate location and orientation from the magnetic field data; our database was previously processed by two independent algorithms: TAPAD (Kroos et al., 2008) and UKF (Chassot, 2009). We use the set processed by UKF after an informal inspection showed UKF trajectories to be smoother. Additional pre-processing methods are discussed APPENDIX D.

### *Generation of STRAIGHT parameters*

The final installation generates the STRAIGHT parameters from the acoustic data. These files are too large to be included on the DVD (you will need approximately 80 GB of free hard drive space). Storing the STRAIGHT parameters offline entails calling the method `saveStraight` with the directory location of the corpus as well as the location to store the STRAIGHT parameters:

```
saveStraight('C:\databases\rgo_ema\',
'C:\databases\rgo_ema\straight\', [125 165]);
```

The final values in this function are the lower and upper bounds for the pitch calculation (use [125 165] for RGO and [80 120] for MAB). This process took approximately 12 hours to complete on a quad-core machine (i5, 4 @ 2.67 GHz) with 4 GB of RAM. Modify the method @utt/getStraight.m to reflect the location of the STRAIGHT parameters.

### *Verifying installation*

Check your installation by opening Matlab and running `checkInstallation.m` in the “C:\ConFAC\examples\” directory. This script executes portions of the required toolboxes in isolation to allow the user to determine where (if any) errors occur. If an error occurs, check your Matlab path to ensure that all necessary directories are included or if there are any conflicts with existing functions of the same name. Proceed to check basic ConFAC functionality by running `demo1.m`, `demo2.m`, and `demo3.m` in the same directory. Successful execution of these scripts almost certainly guarantees ConFAC is installed correctly.

## CONFAC CLASSES

Understanding ConFAC begins by understanding the four ConFAC classes and their relationship to each other. ConFAC was developed as a set of object-oriented classes related by the class diagram in Figure 48. Important properties, methods, and roles of each class are discussed in each sub-section below; a complete description of all class methods can be found using Matlab’s `doc` command.

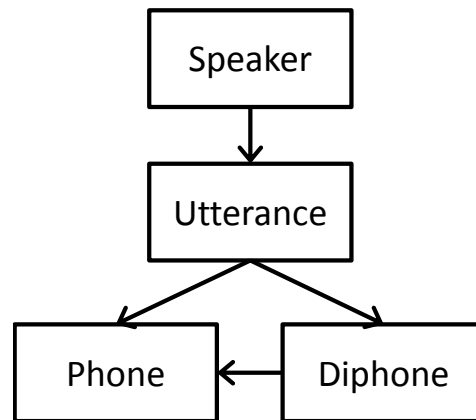


Figure 48 Diagram showing the relationship between the Matlab classes. The speaker class contains an array of utterances, which in turn, contain arrays of phones and diphones.

#### *Speaker class*

The speaker class (@spk) is the highest level class. It represents a collection of utterances and contains methods for loading and searching utterances as well as performing unit selection and synthesis. This class also contains methods for calculating global statistics as well as the main methods for unit selection and synthesis. Its primary property is an array of utterances.

#### *Utterance class*

The Utterance class (@utt) is the primary storage class. It holds raw data (e.g. audio and articulatory), calculated features (e.g. MFCC, Maeda, pitch, loudness), and methods to load the offline STRAIGHT features seamlessly as if they were held in memory. While its main purpose is handling the data, it also contains methods for calculating features and spectrally smoothing the STRAIGHT spectrum. The Utterance class contains arrays of phones and diphones used during unit selection.

Table 11

The properties of the “utterance” class are listed below. This class stores all the raw data, while phone and diphone objects link to a particular utterance.

Property	Sampling rate	Description
wav	16 kHz	acoustic waveform
f0raw	1 kHz	STRAIGHT pitch
ap	1 kHz	STRAIGHT aperiodicity
n3sgram	1 kHz	STRAIGHT spectral envelope
MFCC	1 kHz	13 <sup>th</sup> order Mel-cepstral analysis
artData	200 Hz	6 channel electromagnetic articulography
daf	50 Hz	discrete articulatory features
Maeda	200 Hz	Maeda parameters
txt	–	orthographic transcription
htkWords	–	contents of corresponding *.lab file
phones	–	array of phones
diphones	–	array of diphones

### *Unit class*

This class (`@unit`) is an abstract class representing several types of linguistic units (e.g. phone, diphone, triphone, syllable, word, phrase). The unit class acts a prototype for any derived classes (e.g. `@phone` and `@diphone`). A unit is defined by a start and end time within an utterance. Its main purpose is to provide common methods within to access the utterance-level data. For example, when performing unit selection, every diphone is described by an equal number of features. This is performed using the method `@unit\equalFeatures`. The current approach is to sample continuous features at the beginning, middle, and end of a unit and represent discrete features (e.g. word initial unit, duration) with a single value. Since this representation is independent of time, it is highly recommended to include duration as a feature.

### *Phone class*

The phone class (`@phone`) is derived from the unit class. It is the simplest class and has almost no unique methods. Its properties are basic: start time, end time, and label (e.g. /ah/).

### *Diphone class*

The diphone class (`@diphone`) is also derived from the unit class. Since unit selection is performed at the diphone level, this class is the functional unit of the database. A diphone spans two neighboring phones; its start time must fall within the duration of the left phone and its end time must fall within the duration of the right phone. Times are initialized mid-phone, but are adjustable to improve joins with other diphones. Its label is a dependent property calculated by concatenating the labels of its composite phones. The diphone class contains special methods for estimating diphone similarity and the goodness of two diphones being joined (i.e. How natural is the transition?).

### EXAMPLES

The examples below illustrate both basic (e.g. loading, analyzing, listening) and advanced (e.g. accent conversion, text-to-speech synthesis) functionality of ConFAC.

#### *Loading data*

The speaker class contains methods that load utterances from the database. There are currently two different functions for this purpose: `@spk/populateDB` loads utterances in the order they were recorded, while `@spk/populateMatches` loads only common utterances between RGO and MAB in a specified order. Both functions allow the user to specify the number of utterances to load as well as which features to load (e.g. MFCC, `artData`, discrete articulatory features). During their operation these loaders call, in turn, lower-level loaders for the utterance, phone, and diphone classes. They return an instance of the Speaker class that contains an array of utterances. Utterances can be listened to by issuing the command `@utt/listen` or opened externally in Wavesurfer (Sjölander and Beskow, 2000) using the



command @utt/wavesurfer. Raw EMA data can be viewed with the standard Matlab plot command on the utterance property artData.

```
mab=spk('C:\databases\mab_ema\mat\');
mab.populateMatches(5,10,{'artData'});

u=mab.utterances(1);

u.txt
ans =

'If you want to regulate noise, regulate noise.'

left=u.phones(11).getLeft('artData');
right=u.phones(18).getRight('artData');

u.plotEMA(left,right)
```

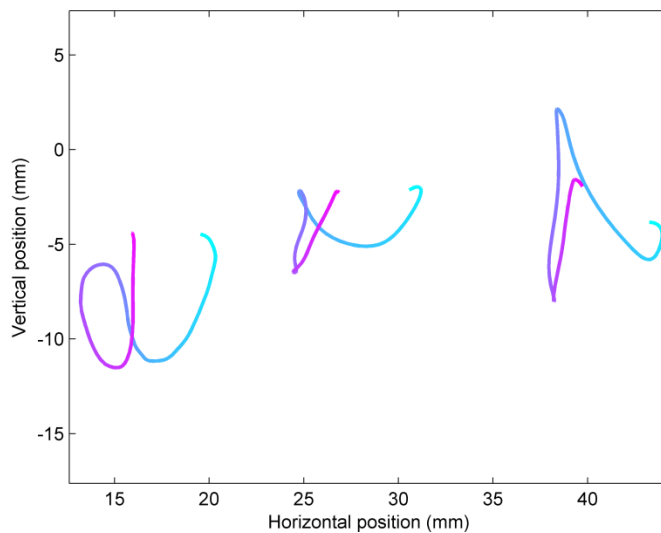


Figure 49 Basic interaction with an utterance object. First a MAB speaker object is created. It loads 5 utterances, starting with the 10<sup>th</sup> utterance in the match list. A handle on the first utterance in the speaker object is created. This utterance contains the sentence “If you want to regulate noise, regulate noise.” The fifth word “regulate” starts with phone 11 and ends with phone 18. A plot of the tongue pellet trajectories is displayed (after zooming) using the command @utt/plotEMA. The far left of the plot is the tongue tip and the far right is the tongue dorsum. Time is displayed as a color shift from blue to pink.

### *Preparing the data*

There are additional functions for preparing the raw data. For the most part, these processes may be called in any order. The only exception is that the function `@spk/fixNaNs`, which estimates missing EMA data, should be called before `@spk/calculateMaeda`. The latter function converts the EMA positional data into the Maeda parameters (see Figure 31) as well as autoscales the Maeda parameters to  $N(0,1)$  using the generic method `@spk/normalizeFeature`. It is also recommended that you autoscale MFCCs to minimize speaker dependencies. The function `@spk/removeOutliers` was implemented to improve the quality of candidate unit selection. It was observed in initial synthesis results that unit selection would sometimes select a candidate that was not a good representation of the target phonetic class (e.g. /ah/). In some instances the unit was mislabeled in the database and in others it was correctly labeled but not typical of that phone. `RemoveOutliers` calculates the mean and variance for each phone in the database using the utterance properties MFCC and duration. In order to balance phonetic diversity with unit integrity, we keep units within two standard deviations of the mean. Figure 50 lists an acceptable way to prepare a speaker object for synthesis; the created speaker object `db` will be used in the remaining examples.

```
load('dbStats.mat','rdb')

db=spk('C:\databases\rgo_ema\mat\');
db.populateDB(100,1,{'artData','daf'});
db.fixNaNs(rdb.NaN);
db.calculateMaeda(rdb.Maeda,1);
db.normalizeFeature('MFCC',rdb.MFCC);
db.removeOutliers(5);
db.initializeIndex;
```

Figure 50 A script demonstrating the standard way to load data and prepare it for accent conversion. It begins by loading a saved structure of speaker dependent statistics (`rdb`). This file was created once using `testNormalize.m` and is used each time data is loaded from the database (there is a similar file for MAB). Refer to each method's help file to understand its input arguments.

### *Saving and loading precomputed databases with .mat files*

The previous two examples are prerequisites for most ConFAC functionality. Since these steps are standard for almost any application, ConFAC includes the ability to save and load a dataset in .mat format. Saving a processed dataset to a \*.mat file and loading it later (rather than re-processing) can save significant time when the datasets involved are large. Save or load any of the custom class variables (i.e. @spk, @utt, @phone, @diphone) using the standard Matlab commands `save` and `load` (see Matlab documentation for more details).

### *Weight training*

Prior to performing synthesis, unit selection target weights must be calculated. This is accomplished through the function `weightTraining4`, which accepts a speaker database and a list of features as input. Weight training assigns a weight to each feature representing its ability to predict acoustic differences between different instances of a particular phone. They are calculated using the weight training method suggested by Hunt and Black (1996). However, multiple linear regression was replaced with partial least squares due to the relatively high dimensionality of our feature space compared to the number of examples in the database (Geladi and Kowalski, 1986). Using the database prepared in Figure 50, we can train target weights by issuing the command:

```
weights=weightTraining4(db, {'MFCC', 'pitch', 'duration'}, 10000, 3);
```

The input arguments specify the database, features, number of diphones to use in training, and the number of times to (evenly) sample each diphone. Here we train weights on standard acoustic features spaced at the beginning, middle, and end (3) of each diphone. 10,000 training examples takes 15 minutes to process on a quad-core machine (i5, 4 @ 2.67 GHz) with 4 GB of RAM. The returned `weights` structure will be used as an input to unit selection.

### *Unit selection*

In addition to the target weights, `weightTraining4` also assigns default values for synthesis parameters. These parameters control various aspects of synthesis, usually representing a tradeoff between two attributes. Table 12 lists the default values and the effect of changing them. In traditional unit selection, units are selected and concatenated from a database to create speech. However, as mentioned previously, ConFAC is a speech *modification* system (e.g. makes a non-native speaker sound more native), which means that some form of the desired utterance is already available. The relatively small size of our database means that some diphones are not well represented. In such cases, the best option is to use the original unit rather than selecting a replacement. This option is enabled by setting `allowSelfUnits` to true.

The operational mode of  $\alpha$  is dependent upon the value of `allowSelfUnits`. When `allowSelfUnits` is false,  $\alpha$  determines the relative importance of concatenation and target costs. On the other hand, when `allowSelfUnits` is true,  $\alpha$  determines the percentage of new units selected from the database costs (see my dissertation for a mathematical formulation of unit selection and the subsequent modifications when `allowSelfUnits` is true). This reformulation of  $\alpha$  has several desirable effects: 1) it replaces units that are furthest away from the target synthesis features first, and 2) it yields a consistent amount of change across utterances and conditions.

The actual unit selection function `@spk/unitSelection6` is not typically called directly by the user, but it is called within common synthesis functions: `TTS`, `leaveOneOut`, and `AC`. Users can control its operation by modifying the parameters stored in the `weights` struct.

Table 12  
Various parameters and their effect on ConFAC synthesis.

parameter	range (default)	low values	high values
$\alpha$	[0,1] (0.33)	selects units that concatenate well	selects units that match the target features
$\beta$	[0, $\infty$ ] (0.33)	selects cut points that concatenate well	selects cut points that meet the desired timing
N	[1, $\infty$ ] (50)	faster, but lower quality synthesis	slower, but higher quality synthesis
allowSelfUnits	[True, false] (false)	(false) Do not consider original units as candidates	(true) Do consider original units as candidates

### *Text-to-speech*

Assuming that a speaker object has been loaded and processed, and target weights have been trained, text-to-speech synthesis can be performed. A crude text-to-speech synthesizer has been implemented for demonstration purposes. Call the speaker function `@spk/TTS` with a speaker object and the desired text. The function will select diphones from utterances in the speaker object to complete the text. In this example we will synthesize the phrase “Hello World.” The function `TTS` first generates a phonetic spelling of the text using the CMU pronunciation dictionary. Diphone target parameters are calculated using average MFCC values from all the examples of that diphone in the database. Lastly, a sequence of diphones is chosen from the database, concatenated, and synthesized.

```

world=db.TTS(weights, 'Hello world. ');
world.phoneList
ans =
    'sil'
    'hh'
    'ah'
    'l'
    'ow'
    'w'
    'er'
    'l'
    'd'
    'sil'

world.listen;

```

Figure 51 A simple text-to-speech example using ConFAC. The dependent property @utt/get.phoneList gives the label of each phone in the utterance. The @utt/listen command plays the synthesized utterance through the computer's speakers.

### *Accent conversion*

Accent conversion is performed with the method @spk/AC. It is the motivation for the entire toolbox and the culmination of 5,000 lines of code. AC is written for maximum flexibility and able to perform accent conversion in many ways. Like SpFAC, the prosodic and segmental conversions can be controlled independently. Furthermore, AC can modify an entire utterance or only specific phones. The minimum input arguments for AC are source utterance, target utterance, and a weight structure. The rest of the parameters will default to perform both the prosodic and segmental conversion for all mispronounced phones and their neighbors. See the built-in documentation to control the other parameters.

```

load('dbStats.mat','rdb','mdb')

rgo=spk('C:\databases\rgo_ema\mat\');
rgo.populateMatches(5,10,{'artData','daf'});
rgo.fixNaNs(rdb.NaN);
rgo.calculateMaeda(rdb.Maeda,1);
rgo.normalizeFeature('MFCC',rdb.MFCC);

mab=spk('C:\databases\mab_ema\mat\');
mab.populateMatches(5,10,{'artData','daf'});
mab.fixNaNs(mdb.NaN);
mab.calculateMaeda(mdb.Maeda,1);
mab.normalizeFeature('MFCC',mdb.MFCC);

mgo=spk('C:\databases\rgo_ema\mat\');
mgo.populateMatches(5,10,{'artData','daf'},...
    'C:\databases\rgo_ema\mat\mab\');
mgo.fixNaNs(rdb.NaN);
mgo.calculateMaeda(rdb.Maeda,1);
mgo.normalizeFeature('MFCC',rdb.MFCC);

load('rgoMFCCpls.mat')
weights.alpha=0.33;
weights.durationWeight=0.33;

newUtt=db.ACspk(rgo.utterances(2),mab.utterances(2)...
    ,[1 1 1],weights,mgo.utterances(2));

```

Figure 52 Example code for performing accent conversion with ConFAC. In addition to the speaker object prepared using code from Figure 50, three more speaker objects must be loaded and prepared (notice that they do not require the function `@spk/removeOutliers` or `@spk/initializeIndex`). The speaker objects `rgo` and `mab` use standard calls to `@spk/populateMatches`. However, the speaker object `mgo` uses an additional argument to the location of the alternate lab directory (refer to Sub-section 6.2). Also, notice that the `rgo` and `mgo` objects are processed with the `rdb` struct while the `mab` object is processed with the `mdb` struct. Weights are loaded and the default values for two of the unit selection parameters are overwritten. Accent conversion is performed using the call on the final line. The binary sequence `[1 0 1]` tells the function to perform the prosodic transform, mispronunciation detection, and the segmental transform.

### *Adding a new speaker*

In the future it may be necessary to add a new database/speaker to the system. If the new speaker is in the same format as RGO or MAB, then the one must: 1) calculate the STRAIGHT files from the acoustic waveform using `saveStraight`, 2) store them to disk, and 3) add the location to `@utt/getStraight`. If the new database is in a different format (e.g. ARCTIC), then it is necessary to define a new class that inherits the `utt` class (e.g. `classdef`

`arcticUtt < utt`). The constructor function for the class should load the new data types into the appropriate properties. For example, the ARCTIC corpus is a simple corpus with \*.wav files holding the acoustic waveforms and \*.lab files for labels. The constructor function should load the acoustic waveform, downsample it to 16kHz, and assign it to the property `@utt/wavData`. Once the primary features are loaded, dependent features (e.g. derivative features) will automatically be computed. ARCTIC label files are not compatible with the `@utt/readLab` function, so a new function must be written. The ARCTIC corpus does not contain any articulatory features, therefore the `@utt/artData` property should be left blank. Finally, STRAIGHT parameters must be calculated using a custom function; refer to `saveStraight` for guidance.

```

classdef arcticUtt < utt

    methods

        function au=arcticUtt(x,fs,labFile)
            %constructor method
            au.wav=resample(x,utt.getFS('wav'),fs);
            au.htkWords=au.readArcticLab(labFile);
            au.MFCC=utt.calculateMFCC(au.wav,utt.getFS('wav'));
        end

        function lab=readArcticLab(labFile)
            %custom function to write
        end

    end
end

```

Figure 53 Adding a different speaker type to ConFAC. The `articUtt` class inherits all of the functions and properties of the regular utterance class. The constructor method uses functions from the utterance class as well as custom functions developed for any new datatypes.



*Adding a new feature*

In the previous sub-section, we looked at adding an ARCTIC speaker that had no articulatory features. If articulatory features were estimated from the acoustic waveform (a process known as articulatory inversion) then they could be saved in the `@utt/artData` property. However, if a database also has true features, it might be desirable to load both types simultaneously. This can be accomplished by adding an additional property to the utterance class (e.g. `@utt/artInvData`), modifying the constructor `@utt/utt` to assign a value to the new property, and adding the new feature's sampling rate to `@utt/getFS`. It can then be accessed like any other property.

Some properties are not continuous. For example, the phone class has binary property to signal when a phone comes at the beginning or end of a word (`wordInitial` and `wordFinal`). These are also dependent properties, which means that they are not stored like normal data, but calculated each time they are needed. To see how they work, look at `@phone/get.wordInitial`).

```

classdef utt < handle

    properties
        artData;
        artInvData;
    end

    methods
        function u=utt(varargin)
            %%%
            %previous utt constructing code
            %%%

            %new code
            u.artInvData=utt.estimateArtFromWav(u.wav);
        end
    end

    methods (Static=true)

        function ai=estimateArtFromWav(wav)
            %%%
            %performs articulatory Inversion
            %%%
        end

        function fs=getFS(type)
            switch type
                case {'wav'}
                    fs=16000;
                case {'artData'}
                    fs=200;
                case {'artInvData'}
                    fs=100;
            end
        end
    end
end
end

```

Figure 54 Modifying the utterance class to add a new feature. In this simple example, we demonstrate: 1) adding the new feature to the properties list, 2) assigning data to it in the constructor, and 3) updating @utt/getFS to record its sampling rate.

## VITA

Daniel Lee Felps received a B.S. degree in computer engineering from Texas A&M, College Station, TX in 2005. He began working on his doctoral studies during the fall semester of 2005 and graduated with this Ph.D. in August 2011. Daniel was awarded the Science, Mathematics and Research for Transformation (SMART) Scholarship from the Department of Defense and served as a summer intern at the National Geospatial Intelligence Agency (NGA) in Washington, DC. He will work for the NGA after graduation.

Daniel may be reached at the Department of Computer Science and Engineering, Texas A&M University, TAMU 3112, College Station, TX 77843-3112. He can also be reached by email at [dlfelps@gmail.com](mailto:dlfelps@gmail.com).