A computational model for studying L1's effect on L2 speech learning

by

Ming Tu

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved August 2018 by the
Graduate Supervisory Committee:

Dr. Visar Berisha, chair
Dr. Julie Liss, member
Dr. Yi Zhou, member

ARIZONA STATE UNIVERSITY

August 2018

ABSTRACT

Much evidence has shown that first language (L1) plays an important role in the formation of L2 phonological system during second language (L2) learning process. This combines with the fact that different L1s have distinct phonological patterns to indicate the diverse L2 speech learning outcomes for speakers from different L1 backgrounds. This dissertation hypothesizes that phonological distances between accented speech and speakers' L1 speech are also correlated with perceived accentedness, and the correlations are negative for some phonological properties. Moreover, contrastive phonological distinctions between L1s and L2 will manifest themselves in the accented speech produced by speaker from these L1s. To test the hypotheses, this study comes up with a computational model to analyze the accented speech properties in both segmental (short-term speech measurements on short-segment or phoneme level) and suprasegmental (long-term speech measurements on word, long-segment, or sentence level) feature space. The benefit of using a computational model is that it enables quantitative analysis of L1's effect on accent in terms of different phonological properties. The core parts of this computational model are feature extraction schemes to extract pronunciation and prosody representation of accented speech based on existing techniques in speech processing field. Correlation analysis on both segmental and suprasegmental feature space is conducted to look into the relationship between acoustic measurements related to L1s and perceived accentedness across several L1s. Multiple regression analysis is employed to investigate how the L1's effect impacts the perception of foreign accent, and how accented speech produced by speakers from different L1s behaves distinctly on segmental and suprasegmental feature spaces. Results unveil the potential application of the methodology in this study to provide quantitative analysis of accented speech, and extend current studies in L2 speech learning theory to large scale. Practically, this study further shows that the compu-

tational model proposed in this study can benefit automatic accentedness evaluation system by adding features related to speakers' L1s.

To my parents, my wife and my lovely daughter

ACKNOWLEDGMENTS

This dissertation would not have been possible without the support of my advisors, committee members, colleagues, family and friends.

Foremost, I would express my sincerest gratitude to my advisors, Dr. Visar Berisha and Dr. Julie Liss for their continuous encouragement and support during my PhD studies. Dr. Visar Berisha has mentored me to design good research ideas and develop solutions to solve them. His openness to diverse thoughts has helped me become a real researcher. Dr. Julie Liss has taught me the correct scientific way of doing research in a new area. They are the best advisors I have ever worked with.

I would also like to thank Dr. Yi Zhou. Thanks for being my committee member and providing helpful advice and inspiring ideas on being a good researcher.

Great thanks to all my labmates during my time at ASU for their assistance and companionship. Thanks to Dr. Beatriz Barragan, Aishwarya Pratap Singh and Anna Grabek for their help in the data collection part of this dissertation. Special thanks to Dr. Alan Wisler, with the companion of whom I spent the first few years in the lab.

The most important part of my PhD study is my wife Yishan Jiao. Having been working in the same labs and with the same supervisors for 7 years, we prove that $1 + 1 > 2$ when a couple studies the same field. Thanks for her companionship in life and inspiration in work.

Thanks to my parents who give as much support as they can to make it possible for me pursuing the commitment throughout my student career. The school teaches me knowledge, and they teach me the rule of life.

This dissertation is also in loving memory of my grandmother who left me one year ago. She is illiterate but her love makes me feel the vitality for life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Problem Statement

Languages are different. Linguistic typology studies classification of world languages depending on their structural and functional features. Because of the diversity of different languages, many criteria can be used to classify languages into different groups (Dryer and Haspelmath, 2013). For example, according to subject-verb-object positioning, languages can be grouped into different sets: SOV (such as French, German, Spanish and Chinese), SVO (such as English and Chinese) and so on, where the abbreviation represents the order of subject(S), verb(V) and object(O). Phonologically, patterns in the structure and distributions of sound systems are investigated by linguistics to classify world languages based on phonological properties. As summarized by Dryer and Haspelmath (2013), properties including vowel and consonant inventories, consonant-vowel ratio, syllable structures, rhythm types, etc. are used to represent the difference in phonology across different languages. Some of those properties mainly measure segmental information while others measure suprasegmental information of one language's phonological system. Those phonological properties result in diverse acoustic characteristics when we listen to speech recordings in d-ifferent languages. One important outcome of the different phonological properties across languages is that when a speaker speaks in a language other than his mother tongue, the speech he produced will be perceived to have accent, which comes from the interplay of the phonological difference of the first language (L1, the speaker's mother tongue) and second language(L2, the language the speaker is speaking). The study in this dissertation will focus on accented speech.

Accented speech is the result of L2 speech being produced by a sensorimotor control system that has overlearned L1 phonological patterns, including both phoneme sound contrasts and prosodic composition. Huge amount of studies have been exploring how a leaner, who is not a L2 native speaker, acquires the phonological patterns of the L2, and where the accent comes from. Among these studies, the most influential theory is the Speech Learning Model (SLM) (Flege, 1995). The SLM hypothesizes that there exists a shared phonological space for both L1 and L2 speech sounds, and uses "equivalence classification" to explain why a learner might not create a new phonetic category for an L2 sound perceived as similar to an L1 sound. Basically, SLM emphasizes the influence of pre-established L1 phonetic categories on the perception of L2 sounds, how it changes over time and also the formation of phonetic categories which is used to produce L2 speech sounds for L2 learners. The SLM mainly focus on the phonetic aspect of a phonological system. Later studies reported that for speech prosody acquisition, the influence from the speaker's L1 also plays a big role in the formation of phonological patterns to produce L2 speech prosody.

Accentedness is usually used to measure the perceived difference between accented speech produced by L2 learners and speech produced by native speakers. There are multiple ways to define accentedness. A more general definition in literature was proposed by McCullough (2013a): accentedness refers to perception of deviations from a pronunciation norm that a listener attributes to the talker not speaking the target language natively. This definition focuses on the difference of foreign accented speech compared to speech produced by native speakers. In second language learning and education practice, accentedness evaluation is very important to designing specific learning targets for different learners based on their level of accentedness, monitoring the learning progress and qualifying or quantifying the learning outcomes. One common experimental design in the study of foreign accent perception is to have participants rate the degree of accentedness in various auditory stimuli, and then to relate these ratings to acoustic properties measured in the stimuli. Much research has been done to study the relationship between perceived foreign accents and acoustic characteristics of accented speech, such as voice onset time (VOT) (Major, 1987a), word duration, stressed or unstressed vowel duration ratio (Shah, 2002), formants movement deviation from L2 acoustic values (Munro, 1993), etc. In addition to the segmental acoustic properties suggested by the findings of previous studies, some studies focus on suprasegmental information, including rhythmic and global temporal properties, of foreign accented speech (Munro *et al.*, 2010; Kang, 2010). Both segmental and suprasegmental acoustic measurements have been shown to be correlated with perceived accentedness. Instead of using human-developed acoustic measurements, a lot of computational models have proposed to automatically extract acoustic measurements and related it to perceived foreign accent. A computational model enables analysis on larger scale with more speakers and provide quantitative analysis of how accentedness can be explained by acoustic measurements. Great success has been reported to use such computational models in automatic accentedness evaluation, and apply them to computer based L2 speech learning and education (Franco *et al.*, 1997; Sangwan and Hansen, 2012; William *et al.*, 2013; Chen and Jang, 2015; Tao *et al.*, 2016; Qian *et al.*, 2017).

Though it is clear that accentedness correlates strongly with how far the phonological patterns of produced accented speech are from patterns of native L2 speech, what has not been studied is whether the distance to the phonological patterns of the speaker's mother tongue matters. According to SLM, phonetic systems of L2 learners respond to L2 sounds by adding new phonetic categories, or modifying existing L1 phonetic categories (Flege, 1995). SLM claims that new phonetic categories may be formed for an L2 sound given sufficient dissimilarity from the closest L1 sound; equivalence classification may block the category formation for an L2 sound, thus the original L1 phonetic category will be used to process both L1 and L2 sound, resulting in similar L2 production with L1 sound. Since SLM mainly focuses on the phonetic system, later studies also investigate the acquisition of L2 rhythm patterns (Rasier and Hiligsmann, 2007; Ordin and Polyanskaya, 2015). The findings in these studies reveal that while the general trend is moving closer to the L2 prosodic patterns as the foreign accent is milder, there still exist effects of L1 rhythm patterns for speakers from different L1 groups. However, existing studies only prove the L1's effect exists in both segmental and suprasegmental properties acquisition, there is no quantitative analysis of the L1's effect. Based on these observations, we may ask:

1. How does the distance from the functional phonological system for accented speech to the actual L1 phonological system relate to the perceived accentedness? If the distance is quantified with acoustic measurements extracted from acoustic signal, are there specific dimensions negatively correlated with the perceived accentedness? Will these L1-related acoustic measurements benefit the acoustic modeling of accentedness perception?

2. Different L1s are at distinctive relative positions with L2 in subspaces of the phonological system, mainly including phonetic space and rhythmic space. For example, German and English are both stress-timed languages, and thus are closer to each other in prosodic space compared to French; Mandarin is syllable-timed language and has different phonetic inventory compared to English, so it is far from English in both subspaces. Will these contrastive phonological properties be transferred to the accented speech during L2 acquisition? Furthermore, will those measurements on phonetic space contribute more to the

perception of accentedness compared to measurements on prosodic space for German speakers speaking English?

To answer these questions, the following hypotheses will be tested in this dissertation:

1. The phonological distance between accented speech and speakers' L1s are also related to perceived accentedness; specifically, L1 related acoustic measurements will have negative correlation with perceived accentedness, considering the negative influence of L1 related factors on L2 acquisition. If these measurements are added to the feature sets for automatic accentedness evaluation, the performance will be improved.

2. Various L1s are relatively different in terms of the distance to L2 in both phonetic subspace and prosodic subspace. Based on this observation, this study hypothesizes that phonological properties in different subspaces (phonetic or prosodic) of accented speech produced by speakers from different L1 backgrounds will have distinct contribution to foreign accent perception. For example, German is close to English in prosodic subspace but relatively far from English in phonetic subspace. With the hypothesis, it can be predicted that prosodic features of accented speech produced by German speakers is less correlated with accentedness score compared to phonetic features.

The above are the research hypotheses this dissertation will test. Specifically, the current study will explore the relationship between acoustic measurements of phonological system with perceived accentedness using a computational model that extracts representative features of both phonetic and prosodic subspaces. The benefit of a computational model lies in its ability to do quantitative analysis on the L1's effect on both phonation and prosody acquisition during L2 speech learning. Given that the deviation of accented speech from the target L2 phonological patterns highly correlates with the accentedness score, this study investigates whether the deviation of accented speech from the original L1 phonological patterns has negative correlation with the accentedness score (i.e. the higher the deviation, the milder the accent); whether integrating L1-related acoustic measurements can improve the modeling capability of accentedness perception, and whether contrastive patterns between L1s and target L2 can be transferred to accented speech.

1.2 Significance of the study

Billions of people are learning a second (or higher order) language nowadays. The number of people living in a second language environment is also increasing with economic globalization. A good understanding of the process of second language learning and accented speech perception is of great importance to successful speech communication in terms of both education, social science and communication science. The current study investigates the perception of accented speech. It aims to achieve better understanding of how the L1s of second language learners affect native L2 speakers' perception of their accentedness phonologically, and how the phonetic system and prosodic patterns contribute to the perception respectively. With a interdisciplinary research methodology combining speech learning and perception theories with speech processing technologies, the current study will have impact on both the theoretic development of second language learning and accented speech perception, and the technologies of Computer Assisted Pronunciation Training (CAPT) and Computer Aided Language Learning (CALL).

1.3 Outline of the dissertation

The dissertation is divided into 8 chapters. Chapter 2 introduces the general background of the current study by reviewing several bodies of research on second language learning and accent speech perception theories and practices: including differential analysis of world languages, second language learning theories, acoustic characteristics of accented speech and computational model of accentedness perception. The motivation and predictions are then presented based on the literature review. Chapter 3 describes the methodology employed in this study including data collection, acoustic analysis and experimental design. Chapter 4 investigates the influence of L1's phonetic system on the accented speech perception. Chapter 5 investigates the influence of L1's prosodic patterns on the accented speech perception, and provide a general discussion on the combined results from this chapter and chapter 4. Chapter 6 combines information of L1's phonetic and prosodic patterns to build a computational model for better automatic accentedness evaluation. Chapter 7 provides a general discussion of the experimental results and tries to extend the current theories on L2 speech learning and accented speech perception. It also introduced the possible implications to both theoretic and practical studies on accented speech. Chapter 8 concludes the current study.

Chapter 2

LITERATURE REVIEW

This chapter will review the literatures related to the research questions in this dissertation. Section 2.1 introduces studies on language typology with a focus on differential analysis of phonological patterns among different languages. The phonological difference between two languages results in foreign accent when learning a second language. Section 2.2 reviews the L2 speech learning theories on both segmental and suprasegmental phonological properties with a focus on the reveal of L1's effect on the formation of foreign accent, which motivates the research questions in this dissertation. Section 2.3 reviews existing studies investigating the relationship between acoustic measurements and perceived accentedness, and studies using computational models to automatically evaluate the accentedness of accented speech. The studies in this section inspires the methodology used in this dissertation. Finally, detailed motivations and expectations of the current study are introduced.

2.1 Differential Analysis of Languages

Language is a system that consists of the development, acquisition, maintenance and use of complex systems of communication, particularly the human ability to do so; and a language is any specific example of such a system. Estimates of the number of human languages in the world varies between 5,000 and 7,000. Languages as communication tools are different in many ways. Recall that when you first learn a second language (L2), the alphabet or words in that language looks so strange, especially for those languages using different sets of characters (for example Chinese and English). You may regard those words as a sequence of graphs without any meanings. Also, when you first listen to a sentence in an L2 or two people talking in an L2, the sound waves are just noise to you. However, you are still aware that those sentences (either in text or sound format) are conveying specific information in a different way from your own language. How are languages different; where do those differences come from?

Language (or linguistic) typology is the science that studies "similarities and differences among languages that do not stem from shared genetic relationship, language contact, or shared environmental conditions" (Moravcsik, 2012b). The goal of language typology is to describe and explain the common properties and structural diversity of the world's languages and how those properties generalize in cross-linguistic case (Bickel, 2001). This discipline includes several subfields, depending on the ways languages are grouped into same classes. An introductory categorization is provided by Moravcsik (2012a):

1. Lexical typology: deals with characteristic ways in which language packages semantic material into words. For example, English uses different words for "foot"/"leg" and "finger"/"toe" while languages like Japanese and Russian use one word to represent "foot"/"leg" ("ashi" in Japanese and "noga" in Russian) and "finger"/"toe" ("yubi" in Japanese and "palec" in Russian).

2. Syntactic typology: deals with characteristic ways in which language packages words into sentences syntactically. For example, according to subject-verb-object positioning, languages can be grouped into different sets: SOV (such as French, German, Spanish and Chinese), SVO (such as English and Chinese) and so on, where the abbreviation represents the order of subject(S), verb(V) and object(O).

3. Morphological typology: deals with characteristics ways in which language form-

s words by combining morphemes. For example, morphological typology categorize languages into analytic languages and synthetic languages. Analytic languages, including Chinese and Vietnamese, contain very little inflection (inflection refers to “the modification of a word to express different grammatical categories such as tense, case, voice, aspect, person, number, gender, and mood”), instead relying on features like word order and auxiliary words to convey meaning while synthetic languages, including most Indo-European languages, form words by affixing a given number of dependent morphemes to a root morpheme and word order is less important for synthetic languages.

4. Phonological typology: dealing with characteristics ways in which sounds are distributed across languages and phonological phenomena such as phoneme inventories, syllable structure, phonological alternations, stress/tone/intonation, prosodic morphology and so on. For example, in terms of consonant-vowel ratio, English is relatively low and Russian is relatively high. In terms of syllable structure, English is relatively complex while Mandarin is relatively simple (Dryer and Haspelmath, 2013).

As introduced in Chapter 1, the current study focuses on phonological patterns of L1 and L2, and how they affect the phonological patterns of accented speech. This section introduces the phonological difference among different languages and will ignore those non-phonological differences. There are several data sources for phonological typology: UCLA Phonological Segment Inventory Database (Maddieson, 1992), Word Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013), URIEL Typological Database (Littel *et al.*, 2016), PHOIBLE (Moran *et al.*, 2014), to name a few. Different databases contain different data sources and language samples. Here, the features provided by WALS are used to illustrate the phonological difference among several languages for the reason that WALS provides simple ways to visualize and download the data.

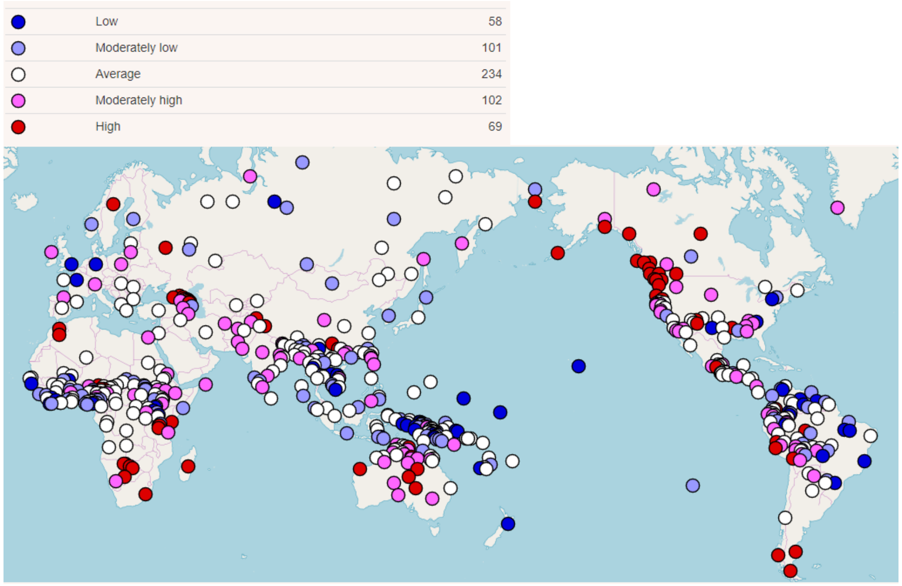


Figure 2.1: Consonant-vowel ratios across sampled languages by WALS. Different colors represent different levels of consonant-vowel ratios: blue for low level, light blue for moderately low level, white for average level, magenta for moderately high level and red for high level.

First, in figure 2.1, the consonant-vowel ratio illustrates how phonological patterns are different across world's languages. Higher ratio means there are more consonants and fewer vowels in that language, while lower ratio means the opposite. Take some commonly used languages as examples: English, German and French all have low ratios; Spanish, Persian and Mandarin have average ratios; Russian has a high ratio.

Next, it is clear to do differential analysis of different languages with phonological patterns, and to illustrate the distances among different languages on phonological feature space. To achieve this, several phonological features pre-summarized by WALS are selected. Based on whether their definitions are segmental or suprasegmental, those features are categorized into phonetic features and rhythmic features. Two

[1] Although Consonant-Vowel Ratio looks like a phonetic feature because it is the ratio of the number of consonants and vowels, most studies regard it as a rhythmic feature (Gil, 1986).

Table 2.1: 19 language phonological features summarized by WALS. The last column indicates whether the feature is phonetic or rhythmic feature. For detailed description of each feature, refer to Dryer and Haspelmath (2013)

Feature name	Phonetic or Rhythmic
Consonant Inventories	Phonetic
Vowel Quality Inventories	Phonetic
Consonant-Vowel Ratio	Rhythmic[1]
Voicing in Plosives and Fricatives	Phonetic
Voicing and Gaps in Plosive Systems	Phonetic
Uvular Consonants	Phonetic
Glottalized Consonants	Phonetic
Lateral Consonants	Phonetic
The Velar Nasal	Phonetic
Vowel Nasalization	Phonetic
Front Rounded Vowels	Phonetic
Syllable Structure	Rhythmic
Tone	Rhythmic
Fixed Stress Locations	Rhythmic
Weight-Sensitive Stress	Rhythmic
Weight Factors in Weight-Sensitive Stress System	Rhythmic
Rhythm Types	Rhythmic
Absence of Common Consonants	Phonetic
Presence of Uncommon Consonants	Phonetic

groups of features represent language phonological patterns on phonetic space and rhythm space respectively. Table 2.1 includes those features' names and indicates whether each feature is phonetic or rhythmic [2] . WALS assigns feature values to languages based on the structural properties of languages that describe one aspect of cross-linguistic diversity. For example, the feature "Rhythm Types" can take five values: Trochaic (left-hand syllable in the foot is strong), Iambic (right-hand syllable in the foot is strong), Dual (system has both trochaic and iambic feet), Undetermined (no clear foot type) and Absent (no rhythmic stress). Those feature values are stored as a number, usually starting from 1, to represent each category they belong to. To visualize those languages on a 2-dimensional space, the numeric values of each feature are employed. If one feature is not applicable to a language, 0 is used instead. As a result, each language will have a 19-dimensional feature vector representing values of those features in table 2.1. Each feature vector is also split into phonetic and rhythmic feature vectors. Since each feature actually indicates a category, to make sure the distances among different categories are the same, one-hot encoding converts the integer feature values to a vector consisting of 0s and 1s. The length of the encoded vector equals to the number of categories that feature can be. For example, the "Rhythm Types" feature has 5 categories. Then, a number of 3 will be encoded as "00010". Multidimensional scaling (MDS), which seeks a low-dimensional representation of those feature vectors in which the distances respect well the distances in the original high-dimensional space, is employed to illustrate the 2-dimensional representation of each language in all phonological feature space (as shown in figure 2.2), phonetic feature only space (as shown in figure 2.3) and rhythmic feature only space (as shown in 2.4) with the encoded language features.

[2] Downloadable from `https://cdstar.shh.mpg.de/bitstreams/EAEA0-7269-77E5-3E10-0/wals_language.csv.zip`

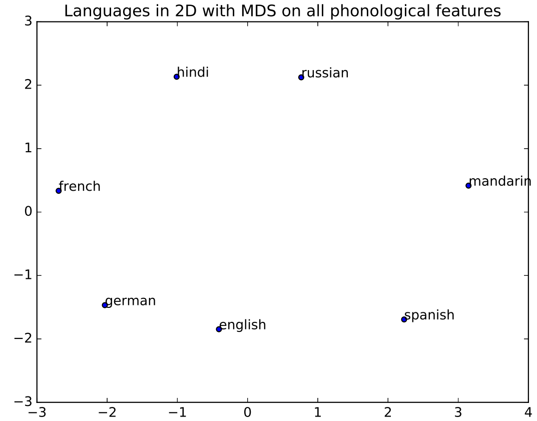


Figure 2.2: 2D visualization of all features with MDS.

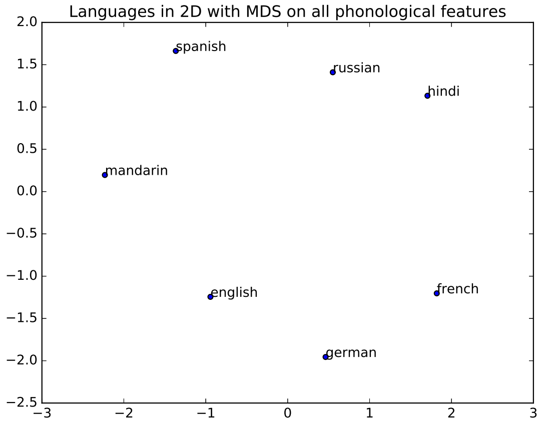


Figure 2.3: 2D visualization of phonetic only features with MDS.

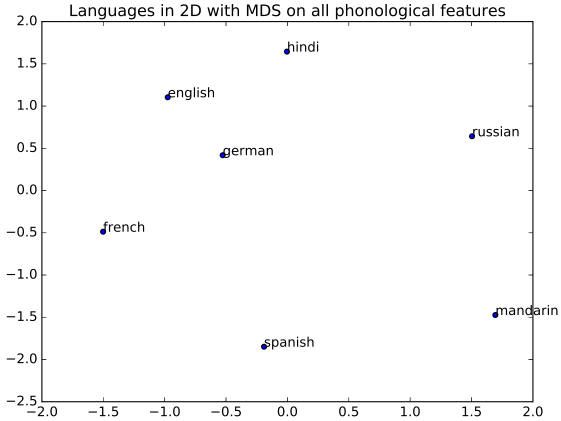


Figure 2.4: 2D visualization of rhythmic only features with MDS.

Table 2.2: Normalized pairwise distance on all features space.

	German	Spanish	French	Russian	Hindi	English	Mandarin
German	0	0.92	0.33	0.67	0.67	0.33	0.83
Spanish		0	0.83	0.67	0.75	0.67	0.50
French			0	0.67	0.58	0.58	1.00
Russian				0	0.42	0.58	0.75
Hindi					0	0.50	0.92
English						0	0.83
Mandarin							0

Along with the 2-dimensional visualization, normalized pair-wise distance matrices are also shown in table 2.2 for all features, table 2.3 for phonetic features and table 2.4 for rhythmic features. The pairwise distance between two languages is calculated as follows: count the number of different values for corresponding dimensions of the feature vectors of two languages without one-hot encoding. This will result in a $N \times N$ matrix where N is the number of languages; normalize the matrix by the maximum value in the matrix. The feature visualization and normalized pair-wise distance matrices suggest that English, German and French are relatively close to each other, while other languages are relatively far from those three languages. Since the scale is different for phonetic and rhythmic feature spaces, it is impossible to compare if a language is closer to English on phonetic feature space or rhythmic feature space. However, within phonetic feature space, it indicates the order by distance to English is $German < Spanish < French \approx Mandarin$; within rhythmic feature space, the order by distance to English is $German < French < Spanish < Mandarin$ (those four languages are used as example). One important question this study wants to investigate is whether those L1 to L2 distance patterns will manifest in the accented speech, and how the relative importance of segmental features and suprasegmental features to foreign accent perception relates to the distance to L2 on phonetic and rhythmic spaces.

The previous features are summarized by linguistics on a high systematic lev-

Table 2.3: Normalized pairwise distance on phonetic features only space.

	German	Spanish	French	Russian	Hindi	English	Mandarin
German	0	1.00	0.29	0.71	0.86	0.43	0.71
Spanish		0	1.00	0.57	0.71	0.57	0.43
French			0	0.71	0.57	0.71	1.00
Russian				0	0.29	0.57	0.71
Hindi					0	0.71	0.86
English						0	0.71
Mandarin							0

Table 2.4: Normalized pairwise distance on rhythmic features only space.

	German	Spanish	French	Russian	Hindi	English	Mandarin
German	0	0.80	0.40	0.60	0.40	0.20	1.00
Spanish		0	0.60	0.80	0.80	0.80	0.60
French			0	0.60	0.60	0.40	1.00
Russian				0	0.60	0.60	0.80
Hindi					0	0.20	1.00
English						0	1.00
Mandarin							0

el. How do those features manifest themselves in the acoustic recordings of different languages? How do the languages' differences manifest themselves in the key parameters of acoustic speech signal, including intensity, pitch, formants, envelop and so on? Several studies have investigated this. An early study (Parmenter and Blanc, 1933) compared the acoustic characteristics between English and French reading speech, and showed that pitch is more important as an element of accent than intensity for French speech, while intensity is more important for English speech. Also, French speech has more pitch variation than English. Studies by Jongman *et al.* (1989); Bradlow (1995); Al-Tamimi and Ferragne (2005) investigated the relationship between vowel inventories and vowel space (defined as the two-dimensional area bounded by lines connecting first and second formant frequency coordinates of vowels (Fant, 1973)), and concluded that vowel space depends on the size of vowel inventory: the larger the inventory, the bigger the acoustic space. The work by Wagner and Braun (2003) showed that predominant factors in voice quality are different across different

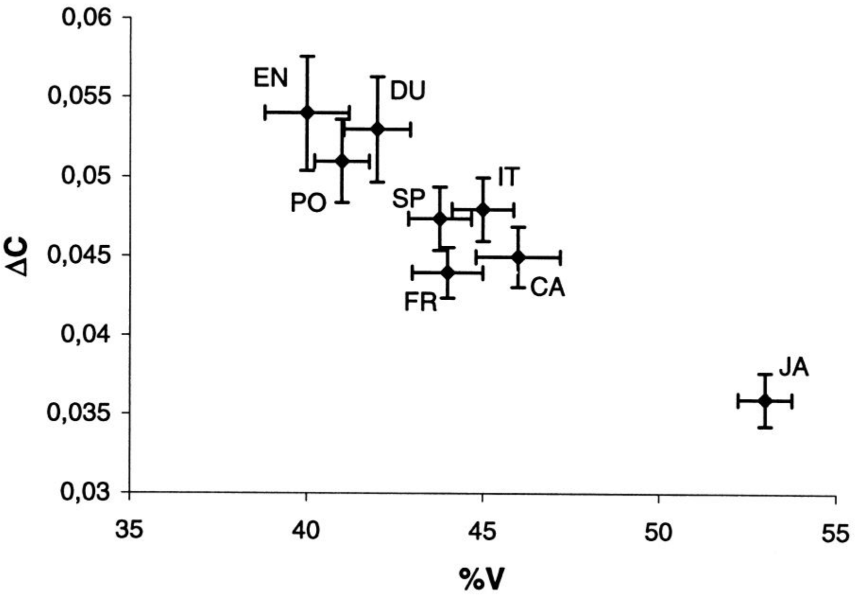


Figure 2.5: Distribution of languages over the ($\%V$, ΔC) plane. EN: English, PO: Polish, DU: Dutch, SP: Spanish, IT: Italian, FR: French, CA: Catalan, JA: Japanese. Taken from (Ramus *et al.*, 1999)

languages. In terms of speech rhythm, an influential study by Ramus *et al.* (1999) investigated the representation of speech rhythm in acoustic speech signal. Several acoustic measurements for speech rhythm are proposed to discriminate the rhythm classes of different languages. Those measurements include the percentage of vocalic segment in an utterance ($\%V$), the standard deviation of consonant intervals (ΔC) and the standard deviation of vowel intervals (ΔV). Figure 2.5 is taken from (Ramus *et al.*, 1999) to show how $\%V$ and ΔC can discriminate languages. Based on this study, other measurements like variational coefficient of consonant/vowel intervals (Dellwo, 2006) and pairwise variability index (PVI) of consonant/vowel intervals (Grabe and Low, 2002) are also proposed. Studies that correlate those linguistically summarized phonological language features with acoustic measurements lay the foundation of the methodology used in this study.

2.2 L2 speech learning theories

In literature, there is a huge body of research on L1 acquisition: how a child acquires a complicated linguistic system including different levels of information without explicit guidance. As summarized by Chang (2010), those studies both investigated the effect of an innately endowed Universal Grammar, and the effect of the timely input on L1 acquisition. A similar research track has been borrowed to study the L2 learning theories: investigating both the influence of already built linguistic system (L1) and some universal effects that are independent of the already built linguistic system. It has been shown that while moving toward to the target L2 linguistic system, L2 learners usually show trackable difference from the implementation of native L2 speakers, which is attributed to the influence of the learner's L1. In terms of phonology, this is where the perceived foreign accent comes from. Considering L1 interference mechanism, i.e., the phonological knowledge transfer from L1 to L2, are focused by the majority of the literature and is more related to the current study, in this chapter research body on L2 speech learning will be reviewed. Specifically, some well-established L2 speech learning theories focusing on the phonetic system acquisition will be introduced first. Then, studies on speech prosody acquisition in L2 speech learning will be covered. The last subsection will focus on the role of L1 in L2 speech learning, and elaborate more on the L1's interference in L2 speech learning.

2.2.1 Phonetic acquisition

There have been studies trying to explain the origin of the foreign accent in producing L2 phonemes. The critical period hypothesis from L1 acquisition was extended to L2 speech learning, positing that there is a critical age or period after which L2 speech production could not be native-like because of the neurological maturation

(Long, 1990). Other studies assume the failure to acquire native-like production of L2 is caused by factors like inaccurate perception of L2 sounds, inadequate phonetic input, insufficient motivation, psychological reasons and incorrect L2 speech learning habit because of incorrect instructions (Flege, 1988). Although all these observations partly show evidence of the origin of the foreign accent, they fail to explain the L2 speech learning process in a systematic way, and how L2 learning is different from L1 acquisition. Nonetheless, there is consensus achieved by the community that the earlier one starts to learn an L2, the better [3] .

Developed by Flege (1995), the speech learning model (SLM) is the most influential study in L2 speech learning literature. Different from the critical period hypothesis, SLM assumes that the phonetic systems used in the production and perception of vowels and consonants is active during the whole life span. It functions like a dynamic system that can encode all phonetic input. As mentioned in (Flege, 1995), "the phonetic systems reorganize in response to sounds encountered in an L2 through the addition of new phonetic categories, or through the modification of old ones". L1 and L2 phonetic categories exist in a shared system, and there is motivation to keep them distinct from each other. This indicates that the formation of the phonetic system of accented speech is based on the L2 learner's already-established L1 phonetic system. To explain this age-related L2 speech learning process, SLM has 4 postulates and 6 hypotheses. The 4 postulates (Flege, 1995) are:

1. The mechanisms and processes used in learning the L1 sound system, including category formation, remain intact over the life span, and can be applied to L2 speech learning.

2. Language-specific aspects of speech sounds are specified in long-term memory

[3] There are still some outliers found by researchers, for example it was reported that both early L2 learners still failed to achieve native-like production while late L2 learners did (Flege, 1995)

representations called phonetic categories.

3. Phonetic categories established in childhood for L1 sounds evolve over the life span to reflect the properties of all L1 or L2 phones identified as a realization of each category.

4. Bilinguals strive to maintain contrast between L1 and L2 phonetic categories, which exist in a common phonological space.

The 6 hypotheses are based on those 4 postulates and on evidence from data analysis in previous studies on speech produced by L2 learners. Next, each hypothesis together with evidence and predicts will be introduced (most of them can be found in the review paper by Flege (1995)).

Hypothesis 1: sounds in the L1 and L2 are related perceptually to one another at a position-sensitive allophonic level, rater than at a more abstract phonemic level. L2 learners will perceive positional allophones in the L2 to the most similar positionally defined allophone in the L1. Studies have shown that it is easier for L2 learners to produce and perceive certain allophones of English phonemes than others. Native Japanese speakers are taken as an example. It is hard for native Japanese speakers producing and perceiving English /l/ and /r/, because in Japanese, there is only one liquid while English has two. Thus, the contrast between /l/ and /r/ is difficult to attain. However, it has been found that the production accuracy of these two liquids depends on phonological environments. In (Strange *et al.*, 1992), the authors showed that native Japanese learners of English characteristically perceive and produce English liquids more accurately in word-final than word-initial position. They attributed this to that the acoustic difference between English /l/and /r/ is more robust in final than initial position (Sheldon and Strange, 1982). This indicates the position-sensitive relationship between L1 and L2 sounds in allpphonic level.

Hypothesis 2: a new phonetic category can be established for an L2 sound that differs phonetically from the closest L1 sound if bilinguals discern at least some of the phonetic differences between the L1 and L2 sounds. The likelihood of the formation of a new phonetic category increases with the dissimilarity between an L2 sound and the closet L1 sound. Several studies have shown that when a novel phoneme (not exists in L1 or very different from L1 phonemes) is encountered, L2 learners can usually produce it accurately. In (Flege, 1987), the authors found that native English speakers can produce the French vowel /y/, a vowel that does not exist in English, relatively accurately compared to native French speakers. Flege (1997) further showed that native Dutch speakers can produce the English vowel /æ/ accurately, and similar results were found in another study on German speakers (Flege and Bohn, 1997). Those findings suggest that if the phonetic differences between the L2 sound to the closet L1 sound are obvious, the production of the L2 sound can be accurate because a new phonetic category is employed to produce the sound.

Hypothesis 3: The likelihood of phonetic differences between L1 and L2 sounds, and between L2 sounds that are noncontrastive in the L1, being discerned decreases as the age of learning increases. For example, the study by Butcher (1978) showed that the perceived distance between /ae/ in English and /ɛ/ in German is greater for German children than adults. Weiher (1975) also showed that German adults, but not children, have difficulty discriminating /ae/ in English and /ɛ/ in German. Based on this hypothesis, it can be predicted that, with the increasing of the age of learning, more sounds in L2 will be inaccurately produced. Thus, a linear relationship between perceived accentedness and age of learning is shown in figure 2.6, in contrast to the sharp discontinuity in the L2 pronunciation ability suggested by the critical period hypothesis (Long, 1990).

Hypothesis 4: Category formation for an L2 sound may be blocked by the

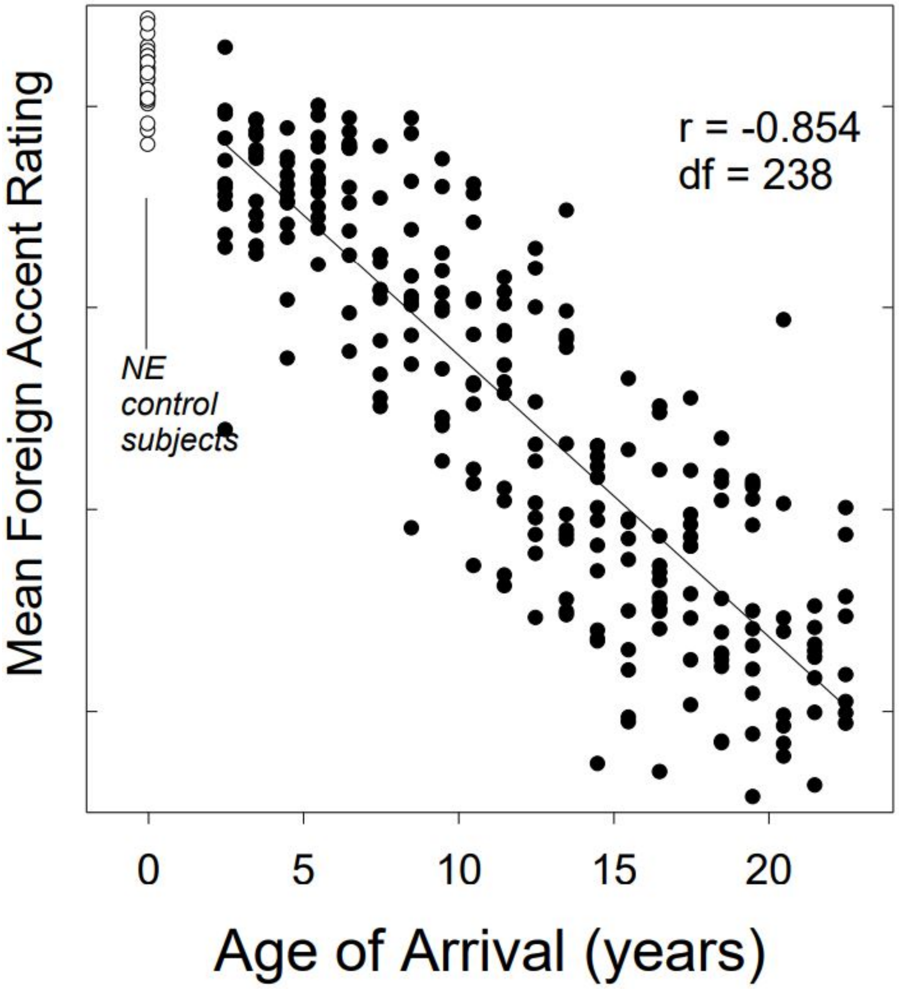


Figure 2.6: The mean foreign accent ratings (Y-axis) of English sentences spoken by native Korean immigrants to US. X-axis represents the age of arrival. Taken from (Flege *et al.*, 1999).

mechanism of equivalence classification. When the block occurs, speakers tend to use a single phonetic category to process perceptually similar L1 and L2 sounds, resulting in inaccurate production of L2 sounds. The study by Flege (1987) showed that French learners who are native American English speakers produce the French phoneme /u/ with second formant (F_2) values higher than native French speakers, which is influenced by the high-F_2 /u/ in English. Chang *et al.* (2008) also reported that native American English speakers also produce Mandarin phoneme /u/ with higher F_2. Flege (1987) further showed that native English speakers produce French voiceless stops with too long voice onset times (VOTs), under influence from the long-lag VOT of English voiceless stops.

Hypothesis 5: The phonetic category established for L2 sounds by a bilingual may differ from a monolingual's if: 1) the bilingual's category deviates from an L2 category to maintain phonetic contrast between categories in a common L1-L2 phonological space; or 2) the bilingual's representation is based on different features, or feature weights, than a monolingual's. The evidence can be found in the study by Munro (1993), where the authors showed that even experienced L2 English speakers, who are native Arabic speakers, produce vowels that are considered to have accent. According to the study, the accentedness was due to non-native production of duration differences between tense and lax English vowels. They suggest that in this case, the L2 tense and lax categories might have been interpreted as long and short categories, which exist in Arabic. The evidence of the second point is shown in the study by Munro *et al.* (1996). This study showed that English learners with Italian as the native language can not produce accurate phoneme /ɚ/, although those learners started to speak English at ten years of age and were rated to have a very mild foreign accent. The authors considered the reason to be related to the retroflex feature that is used to discriminate from other English vowels, but the feature does not exist in Italian.

Hypothesis 6: The production of a sound eventually corresponds to the properties represented in its phonetic category representation. This hypothesis can be regarded as the result of hypothesis 2, 4 and 5, stating that the L2 sound will eventually be produced as specified in phonetic category representation. If the presentation matches the category for native L2 speakers, then the L2 sound can be produced accurately; if new phonetic category for L2 sounds is not formed or different from monolingual's, there will be inaccurate pronunciation.

To summarize, SLM claims that the age of learning has significant influence on second language learning: this can be seen from those hypotheses that the age of learning directly influences the formation of phonetic categories to produce L2 sounds. Also, pre-established L1 phonetic categories will affect the way L2 sounds are perceived, and thus will also influence the formation of phonetic categories for L2 sounds. If sounds in L2 are too close to sounds in L1, then equivalence classification will use the same phonetic category to produce the similar sounds, resulting in perceivable inaccurate pronunciation. Sometimes, phonetic categories built for novel L2 sounds can still be different from native's due to the dissimilation occurs between L1 and L2 phonetic categories to maintain phonetic contrast between categories in a common L1-L2 phonological space. This study mainly reviews the SLM model because it is highly related with the current study in a way that it directly explains how L2 learners develop inaccurate pronunciation of L2 sounds. Another well-known model, the Perception Assimilation Model (PAM) (Best, 1995; Best and Tyler, 2007) mainly deals with how a listener perceptually assimilates contrastive information between his L1 and a new language he does not know or just starts learning. However, this study mainly deals with speakers who are not beginners of L2 speech learning. Thus, the literature on PAM is not reviewed here.

SLM mainly deals with phonetic acquisition, i.e. the segmental inaccuracy of L2 production. However, several studies have shown that inaccurate suprasegmental productions can also result in perceivable foreign accent (Rognoni and Busà, 2013; Winters and O'Brien, 2013). In the next subsection, speech prosody acquisition in the literature will be reviewed to reveal the mechanism L2 learners use to learn the L2 speech prosody.

2.2.2 Prosody acquisition

In the previous subsection, fundamental studies on L2 speech learning are reviewed. Those studies mainly focus on the phonetic part of the whole phonological system. Although the study by Munro (1993) investigated the durations of tense and lax English vowels produced by native Arabic speakers and durations of vowels are related to speech prosody (Ramus *et al.*, 1999), most analysis in these studies only dealt with pronunciation of specific phonemes; some even used isolated phonemes or words (Flege, 1987). Whether the theories proposed by these studies can be applied to prosodic inaccuracy of non-native L2 speech is still questionable (Rasier and Hiligsmann, 2007). On the other hand, a review survey by Gut (2009) showed that for all studies on L2 speech learning from 1969 to 2008, L2 intonation was only investigated in nine studies and L2 speech rhythm was only investigated in four studies (Mennen, 2004; Altmann, 2006; Rasier and Hiligsmann, 2007; Lin and Wang, 2008). This indicates that the speech prosody in L2 speech is quite underexplored. This subsection will review literatures on acquisition of speech prosody acquisition during L2 speech learning.

The study by Mennen (2004) investigated how the non-native speakers of Greek whose L1 is Dutch realize the timing of a phonologically identical rise: nonfinal or prenuclear rises. This phonological property was realized differently by native Dutch and native Greek speakers: 1) at a different time: the peak in the rise appeared earlier in Dutch than in Greek. 2) The peak time in Dutch depends on the the phonological length of the vowel of accented syllable while Greek did not. By analyzing the timing patterns of the rise using five native Dutch speakers speaking Greek, the authors concluded that there existed a bi-directional interference in the realization of the rising accent: the L1 Dutch affected the realization in Greek and the L2 Greek also affected the realization in Dutch. The dissertation by Altmann (2006) studied the perception and production of advanced learners of English with different L1 backgrounds (Arabic, Chinese, French, Japanese, Korean, Spanish, Turkish) to investigate the effect of L1 stress properties on the L2 acquisition of primary word stress. The results showed that native speakers of L1s with predictable stress found it difficult to locate the stress in English although they were able to produce the correct stress patterns; native speakers of L1s without word-level stress or predictable stress performed well in stress perception but had difficulties in stress production. These results seem to contradict the SLM: good perception of stress patterns does not mean good production. Rasier and Hiligsmann (2007) reviewed the prosody acquisition of L1 learning and proposed a general framework to study the prosody transfer from L1 to L2 in L2 speech learning. The model was tested in a study of accent in L2 Dutch proposed by native French speakers and L2 French produced by native Dutch speakers. Their results showed that the difference between French and Dutch on accent placement influenced the acquisition process of accentuation. The "Markedness" proposed in Eckman (1977) is an important factor in predicting and explaining learning difficulties in L2 prosody learning.

The previous studies mainly focus on stress and accent. The following introduced studies in this paragraph investigate the rhythmic properties' acquisition in terms of duration and duration variability measurements, which have been shown to able to discriminate among languages belonging to different rhythmic classes Ramus *et al.* (1999); Grabe and Low (2002). Those measurements include:

1. ΔV: the standard deviation of vocalic intervals

2. ΔC: the standard deviation of consonantal intervals

3. $\%V$: percentage of vocalic intervals in the sentence

4. $VarcoV$: the standard deviation of vocalic intervals divided by the mean vocalic interval duration and multiplied by 100

5. $VarcoC$: the standard deviation of consonantal intervals divided by the mean consonantal interval duration and multiplied by 100

6. $nPVI-V$: the normalized PVI for vocalic intervals

7. $rPVI-C$: the raw PVI for consonantal intervals

One study (Stockmal *et al.*, 2005) examined speech rhythm of the Latvian produced by native Russian learners. In their result, there was no clear increase in vocalic variability between experienced and low-level learners, despite the fact that Latvian is significantly less stress-timed than Russian. They concluded that even if the learner's L1 is stress-timed and has higher vocalic variability, at the early stage of acquisition the accented speech can still match the L2 in terms of lower vocalic variability. They also found that the consonantal duration variability increased significantly during L2 acquisition and attributed to the difficulties of consonants articulation. White and Mattys (2007) used all the seven rhythmic measurements, showing that those measurements were able to separate stress-timed English and Dutch and syllable-timed Spanish and French. They also applied the measurement to quantifying the influence of L1 on L2 rhythm acquisition when switching between stress-timed and syllable-timed. In an experiment consisting of native Spanish subjective speaking English and native English speakers speaking Spanish, it was found that the $VarcoV$, $nPVI-V$ and $rPVI-C$ were in the intermediate stage during the transfer from L1 to L2, indicating clearly the influence of L1 rhythm on l2. In the experiment consisting of native Dutch subjective speaking English and native English speakers speaking Dutch (both the two languages are stress-timed), it was found that there was no clear influence of L1. The authors believed that if L1 and L2 are already rhythmically similar, the L2 learners tend to make little accommodation and use their L1 rhythmic patterns. Lin and Wang (2008) examined the accented English speech produced by native Mandarin speakers in terms of four measurements of speech rhythm: $\%V$, ΔC, $rPVI - C$ and $nPVI - V$. With the reading and conversational recordings of 6 subjects, the authors showed that the value of $\%V$ of Mandarin accented English is in the middle of the value of native English speakers (lower) and native Mandarin speakers (higher). They explained that this indicated the L1 rhythm patterns had an effect on L2 rhythm patterns in terms of $\%V$. The average nPVI value was very close to native English speakers, and the authors attributed this to that those Mandarin subjects mastered the vocalic variability. However, the average values of the other two measurements are way higher than native English speakers. The authors believed it was because the consonantal duration patterns were much harder to acquire for Mandarin speakers when speaking English. Similar results were also reported by Kawase *et al.* (2016). In this study, the rhythmic acquisition of native Japanese (mora-timed) learners of English (stress-timed) was studied. Li and Post (2014) conducted experiments on durational variation in L2 English productions by L1 Mandarin learners and L1 German learners and compared it to native control values in English. The results showed that the L1 groups followed comparable developmental paths in their acquisition of vocalic variability and accentual lengthening. However, the two L1 groups diverged in the proportion of vocalic materials in their L2 utterances and indicated L2 acquisition patterns that are consistent with direct transfer from the L1. Thus, they claimed that there was a multisystemic model of L2 rhythm acquisition. Both transferred L1 knowledge and universal effects independent of L1 played a role. Ordin and Polyanskaya (2015) did similar experiments to examine the differences in durational variability (several rhythmic measurements) between proficiency levels in L2

English spoken by French and German learners. They found that speech rhythm in L2 English learners in both groups developed from more syllable-timed toward more stress-timed patterns irrespective of the L1 had similar rhythmic patterns. However, they also showed that there were differences between the German and French groups: German learners achieved a degree of durational variability typical of the target language while French learners exhibited lower variability than native speakers.

Some recent studies investigated the relative importance of suprasegmental measurements to accentedness perception compared to segmental measurements. Rognoni and Busà (2013); Winters and O'Brien (2013) transplanted the prosody measurements (F0 and duration) between native English speech and accented English speech in both directions to analyze the relative importance of segmental and suprasegmental features' contribution to accentedness perception. They both found that though prosodic features contributed to the perception of accentedness, segmental features were more important than suprasegmental features. The study by Polyanskaya *et al.* (2016) applied the similar transplantation method to speaking rate and speech rhythm and concluded that speech rhythm contributed more to accentedness perception than speaking rate. A later study by van Maastricht *et al.* (2017) further investigated the interplay of different prosodic measurements including intonation, rhythm and speech rate. The authors found that while all measurements contributed to accentedness perception, intonation contributed the most for Dutch learners. However, all of these studies only did the transplantation on one foreign language (Italian, German, French or Spanish) and the contrastive information among different L1s was ignored. Another work by Saito *et al.* (2016) studied the relative contribution of segmental and suprasegmental to accentedness at different proficiency levels through regression analysis. Their subjects were Japanese who were learning English at different stages.

The prosody acquisition during L2 speech learning can be summarized as following:

1. Although there are evidences showing that some universal effects exist in prosody acquisition, most studies report the influence of the L1 on the prosody production of L2. This is similar to the phonetic acquisition.

2. Not all of the prosodic properties depend on the L1 during speech prosody acquisition, despite the fact that those properties can well discriminate between L1 and L2.

3. When the contrastive information between L1 and L2 can be well perceived, the prosody acquisition follows a path from L1 prosody features to L2 prosody features; when the contrastive information between L1 and L2 is not well perceived or the L1 and L2 prosodic patterns are very close, there is no clear sign of the influence of L1 prosodic patterns.

2.2.3 Role of L1 in L2 speech learning

In the last two subsections, studies dealing with the phonological system acquisition during L2 speech learning are introduced. However, many studies only investigate one pair of L1 and L2: a one to one mapping, which can not reveal whether the difference of L1s can be projected to the accented L2 speech. Combining the acquisition of L2 in both segmental and suprasegmental perspective, it can be found that different L1s can result in different developments of L2 acquisition. A following question is how the segmental and suprasegmental production developments of L2 learners from different L1s are different. Since L1s have very different segmental and suprasegmental characteristics compared to L2, L2 learners from those L1s should undergo different procedures in both segmental and suprasegmental feature space, as in the findings by Ordin and Polyanskaya (2015), although there exist some universal effects. In this subsection, a brief introduction of studies examining multiple L1s and one or multiple L2s are reviewed. Arslan and Hansen (1997) calculated four temporal measurements: word-final stop closure duration, VOT, average voicing duration and word duration, across three L1 accents (German, Mandarin and Turkey) for a set of English words ("target", "teeth", "catch", "communication"), which included a stop consonant in the initial position. While word-final stop closure duration was found to be most discriminative among accents, various accents showed very different patterns in terms of the four measurements. English words with Mandarin accents are the most different from native produced words. German and Turkey speakers are relatively closer to native produced words compared to Mandarin. McCullough (2013b) investigated the correlation between different segmental measurements in non-native speech and the perceived accentedness. Speakers from different L1 backgrounds (Hindi, Mandarin and Korean) were rated and analyzed based on their produced English speech. They showed that Hindi had the strongest accent compared to Mandarin and Korean speakers. Analysis of measurements including VOT, vowel quality (measured by F1 and F2 of vowel), vowel durations and F0 differences indicated that non-native speech produced by Hindi speakers showed clear difference compared to Mandarin and Korean speakers, while Mandarin and Korean speakers had similar patterns. For suprasegmental measurements, Ramus *et al.* (1999) studied the rhythmic properties across eight languages (English, Polish, Dutch, French, Spanish, Italian, Catalan and Japanese) and applied acoustical rhythmic measurements to language discrimination. They plotted these eight languages on a three dimensional rhythmic space consisting of 1) the proportion of vocal intervals within the sentence 2) the standard deviation of the duration of vocalic intervals within each sentence 3) the standard deviation of the duration of consonantal intervals within each sentence. The results indicated that there may be more information decided by speech rhythm rather than just the classification of stress-, syllable and mora-timed languages. Also, the difference among different L1s was very obvious. This study inspired the work by White and Mattys (2007) and Lai *et al.* (2013), where the authors applied similar acoustic analysis of the rhythm properties of both reading and spontaneous L2 speech. White and Mattys (2007) applied similar acoustical rhythmic measurements to quantifying the influence of L1 on L2 rhythm. They expected that speakers switching "rhythm class (stress-timed or syllable timed)" should show rhythm scores different from both their native and target languages. They found that the standard deviation of vocalic interval duration divided by the mean vocalic interval duration offered the most discriminative ability of L1, L2 and L1 accented L2, which suggested L1 accented L2 is at an intermediate stage during the transfer from L1 to L2, and speakers with different L1 backgrounds show differences in their accented L2 speech in terms of these rhythmic features. While previous studies used reading speech as material, the work by Lai *et al.* (2013) investigates the rhythmic measurements of spontaneous L2 speech produced by speakers from different L1 backgrounds. TOEFL Practice Online assessment of 239 speakers from 50 L1 backgrounds was used as speech material. They compared the rhythmic properties of accented L2 speech with measurements proposed in the study by Ramus *et al.* (1999), and showed the difference between rhythmic properties of L1 speech and L1 accented L2 speech, as well as the difference between rhythmic properties of reading and spontaneous accented L2 speech. However, the different rhythmic properties of different L1s were mostly kept in the L1 accented L2 speech.

It can be concluded that in the phonological space of languages, at least on some dimensions (including both segmental and suprasegmental dimensions), L2 learners from different L1 backgrounds follow a speech acquisition path that starts from their L1s and goes towards the target L2. On the path, the L2 speech produced by different

L1 learners still show distinctions that depend on the L1s. However, the methodologies used in these studies are only capable of showing the existence of L1's influence on the formation of accent and different phonological characteristics of accented speech by speakers from different L1s, but are very limited to quantify the L1's effect on L2 speech learning.

2.3 Computational models for accentedness perception

Previous sections reviewed the language differences, second language acquisition of both segmental and suprasegmental phonological properties and how different L1s will result in different development pathes in L2 speech learning. This section deals with the learning outcome: accentedness, which is usually defined as the degree of perceived foreign accent. Specifically, this section focuses on how accentedness is related to acoustic characteristics of accented speech, and whether the accentedness of a speaker can be predicted with computational models given produced accented speech. Furthermore, investigating the relationship between perceived accentedness and acoustic measurements is also a commonly used methodology in L2 speech learning studies (Ordin and Polyanskaya, 2015; Saito *et al.*, 2016).

Abundant studies have been done to investigate the relationship between perceived accentedness and acoustic information, such as segmental and suprasegmental measurements, and these studies lay the foundation of computational models for accentedness perception. Segmental features measured in short time periods, including voice onset time (VOT, defined as the duration between the release of a consonant and the onset of voicing)(Lisker and Abramson, 1964; McCullough, 2013b,b), pronunciation of vowels and consonants (Flege, 1995; Deterding, 2006; Sangwan and Hansen, 2012), vowel quality, vowel duration, short-time F0 and harmonics (McCullough, 2013a,b), have been shown to contribute significantly to the perception of accentedness. Suprasegmental measurements are also found to be significantly important to accentedness perception. For example, Hardman (2014) investigated the interlanguage match effect of Mandarin-accented English. They found that Mandarin accent had a large negative effect on intelligibility, but the talker's accuracy was still high. They considered that low intelligibility was due to a combination of the segmental variation and its misalignment with higher levels of prosody. This means that accented speech can be segmentally close to native speech, but still results in low intelligibility and high accentedness score due to suprasegmental mismatch. The studies by Munro and Derwing (2001); Mok and Dellwo (2008); Kang (2010) found that suprasegmental measurements such as speaking rate, consonantal/vocalic/syllabic durations, pauses, stress and pitch range of non-native L2 speakers also contribute to the perception of accentedness. How to convert those measurements (although some of them are computed automatically, most are measured with human labor) in previously introduced studies to acoustic features that can be computed automatically from acoustic signal is the main goal of a computational model for accentedness perception.

Those studies in phonological linguistics have inspired research on computational models for accentedness perception. In the field of computer-assisted pronunciation training (CAPT) and computer-aided language learning (CALL), which has proved to be able to improve language learning, especially word pronunciation (Neri *et al.*, 2008), many studies investigated improving second language learning and education using computer based accentedness evaluation systems. The goal of automatic accentedness evaluation is to build a statistical machine learning model that predicts the accentedness score of non-native speakers, which is supposed to be highly correlated with humans' ratings of accentedness. Acoustic features that can represent the segmental and suprasegmental measurements of accentedness speech are extracted in an automatic way and the evaluation model is responsible for learning the mapping from acoustic features to accentedness score in a supervised learning way. Some studies only focus on the pronunciation part of non-native L2 speech, while some recent work also includes suprasegmental features.

The first work that aims to develop computer based systems for language learning instruction was conducted in Speech Technology and Research Laboratory at SRI International. Their early pronunciation scoring systems (Bernstein *et al.*, 1990) were designed as text-dependent, which means nonnative speakers must read fixed words or sentences. Text-dependency makes these systems very hard to use for real language training and evaluation. Their following work focused on a text-independent system. The corpus the authors developed consisted of 100 native French speakers from Paris and 100 American students speaking French. Nonnative French speakers were asked to read designed speech materials including common sentences, newspaper sentences and imitated speech after listening to a native reading the same sentence. Nonnative speakers were rated by language experts on a 1-5 (unintelligible to native quality) scale. The task was to automatically grade the pronunciation performance of nonnative speakers. In the study by Neumeyer *et al.* (1996), an automatic pronunciation scoring system was proposed based on an ASR system. First, nonnative speech was segmented using the alignments provided by the ASR system. Four scores were calculated including Hidden Markov Model (HMM) log-likelihood score on each segment, phone classification scores on each segment, segment duration scores calculated by log probability of a phone duration model trained with native speakers, and time scores calculated by averaged and normalized time between syllables. Correlation with human raters showed that segment duration scores provided the highest correlation (sentence level: 0.46, speaker level: 0.74).They reported improvement in their following work. For example, sentence level correlation improved to 0.50 and speaker level correlation improved to 0.88 by using average phone segment posterior probabilities, which was calculated by frame-based phone posterior probability. And using score combination (input to linear or nonlinear regression models), sentence level correlation rose to 0.62 (Franco *et al.*, 1997). This line of research was extended to assessing pronunciation quality on individual phone segment using the same database (Kim *et al.*, 1997). Listeners were asked to only rate certain segments, and same scores were calculated on each segment. Similarly, log-posterior probability scores provided the highest correlation. The overall speaker level correlation was 0.88. The authors show that human-machine correlation was higher than human-human correlation on both phone segment level and speaker level. Xi *et al.* (2010) reported a summarization and extension of their previous work on pronunciation scoring. In this research, Spanish was the L2 speech and Spanish learners were native American English speakers. They found word duration scores provided better results than phone duration scores.

Sangwan and Hansen (2012) proposed an automatic accent analysis system of Mandarin accented English using phonological features. With a trained HMM-based phonological feature classification system, they built two Markov Models to capture the dynamics of phonological features for both American English speech and Mandarin-accented English. State transitions and state durations of phonological features were believed to carry very important accent-related information. For a given English word produced by a Mandarin speaker, accentedness was represented by delta log-likelihood that is calculated by the log-likelihood of the two trained phonological features Markov models. The accentedness indicator was on a scale from -1 to +1 (from non-native like to native like). Through experiments on CU-Accent corpus (Angkititrakul and Hansen, 2006), a correlation of 0.8 was reported between human assigned scores and scores provided by the proposed system. William *et al.* (2013) proposed a new algorithm for automatic accentedness evaluation. The system had two parts. In the alignment part, speech utterance was processed using a Weighted Finite State Transducer (WFST) based decoder of an ASR system to automatically estimate the pronunciation mismatches including substitution, deletion and insertion errors. In the scoring part, two scoring systems which utilized the pronunciation mismatches from the alignment phase were proposed: a WFST-scoring system to measure the degree of accentedness on a scale from -1 (non-native) to +1 (native), and a Maximum Entropy (ME) based system to assign perceptually motivated scores to pronunciation mismatches. The proposed algorithm was also evaluated on CU-Accent corpus. The results showed that the correlation between human raters and machine system was as high as 0.89. Chen and Jang (2015) proposed a learning-to-rank based automatic pronunciation scoring framework. The motivation was that they believed it is easier for a human rater to make a relative judgement than to assign an exact score. The authors used similar feature sets as the study by Kim *et al.* (1997). These phone-level scores were then converted to word-level scores, which were used to train the learn-to-rank model. The output of the learn-to-rank model was quantized onto the 1 (unintelligible)-5 (intelligible) scale, which was the rating scale for listeners. The results on a Taiwan Mandarin speech corpus showed that the proposed system achieved a better correlation compared to human ratings. Rasipuram *et al.* (2015) developed an automatic acccentedness evaluation system based on comparison of instances of native and nonnative speakers at the acoustic-phonetic level. The main advantage of their system was its capability to go beyond the instantaneous phoneme level scoring, and provided utterance level and speaker level scoring of accentedness. A Deep Neural Network (DNN) based acoustic model was used to map the input feature vectors into sequences of HMM states. A dynamic programming based sequence matching algorithm was employed to calculate the pronunciation mismatch between nonnative speakers and native speakers. Human ratings on a scale of 0 (no foreign accent)-6 (foreign accent) was collected for Finnish, German and Mandarin-accented English and final reported correlations for Mandarin-accented English between human raters and system's output were 0.66 on the sentence level and 0.73 on the speaker level.

Nativeness evaluation of nonnative English speakers was also introduced into Interspeech 2015 paralinguistic challenge (Schuller *et al.*, 2015). The dataset included nonnative English speakers with multiple mother tongues including Mandarin. In their baseline system, Opensmile (Eyben *et al.*, 2010) was used to extract acoustic features from utterances. Support vector regression (SVR) was employed to predict the nativeness score. The challenges of this task were that the rating scale of train, development and test sets were different and it was a cross-corpora task. The reported correlation coefficient between predicted nativeness and human ratings was around 0.4 on sentence level. Several papers were submitted to improve the baseline system. In the study by Grósz *et al.* (2015), instead of using SVR, DNN and Gaussian Process regression were employed for regression analysis with the same acoustic feature sets as the baseline system. They reported higher correlation coefficients than the baseline system. Ribeiro *et al.* (2015) developed several feature sets besides the baseline features. Their feature sets, including phonotactic models (for language identification) based features, n-grams counts based features and ivectors, were both employed as the input feature sets for SVR, which means three complex models needed to be prepared: a language identification model, an ASR model and an ivector extraction model. The correlation reported on test set was 0.58, which was much higher than the baseline system. Black *et al.* (2015) also focused on feature development for nativeness evaluation. Different from previous study that employed several feature sets from other related tasks, this paper developed multiple feature sets at multiple time scales to include both segmental and suprasegmental information. These feature sets consisted of data-driven features, including baseline acoustic features and other low level descriptors used in their previous studies, and knowledge based features, including utterance level pausing features, speaking rate related features, lexical stress, intonation and speech rhythm related features, and phone-level pronunciation features. Extraction of knowledge based features needed an ASR system trained on native speakers to provide alignment and phone-level likelihood. Their result was the best among all submissions, with correlation coefficient as high as 0.75 on test set.

In recent studies, state-of-the art ASR systems based on recent advancement in DNNs are investigated in automatic non-native speech assessment. Tao *et al.* (2016) trained a non-native spontaneous speech ASR system, using over 800 hours of native-speech recordings. They investigated three ASR systems: a traditional GMM-HMM system, a DNN-HMM system and a GMM-HMM system using DNN as feature extractor. Several feature sets for nativeness evaluation, part of which was based on the trained ASR systems, were extracted from non-native speech. These feature sets were categorized into fluency, rhythm/intonation/stress, pronunciation, grammar and vocabulary use of the non-native speech, covering both the segmental and suprasegmental measurements of non-native speech. Their system could achieve as high as a 0.78 correlation coefficient with human raters on non-native spontaneous speech. In the study by Qian *et al.* (2017), a Recurrent Neural Network (RNN) acoustic model was applied to children's speech recognition to improve the automatic assessment system of children's non-native speech. Their motivation was that most current automatic accentedness assessment systems used ASR trained on adults which did not perform well for children's speech. Their ASR system was purely trained on children's speech, and the same feature sets as in the study by Tao *et al.* (2016) were used to represent the proficiency of children's speech. Their final reported correlation coefficient between the system's prediction and human raters was 0.76 on non-native children's speech.

To summarize, the most important part of computational models for accentedness perception is the feature extraction, i.e. to extract foreign accent related representations from acoustic signals. Most studies use ASR systems trained on native L2 corpus to quantify how well the L2 learners pronounce each segment (phoneme or word), and results have shown that those measurements based on ASR can give good results. Some small scale studies also prove the effectiveness of features such as VOT, formants, pitch, and so on. For speech prosody, no standard feature extraction scheme exists yet. Recent studies extract durational measurements based on computer-generated phoneme forced-alignments to represent speech prosody, and good results have been reported. The most beneficial part of computational model is its ability to quantify the relationship between a specific phonological properties and accentedness. This could facilitate better understanding of L1's effect on L2 speech learning, thus motivating the methodology in this dissertation.

2.4 Motivations and predictions

1. Clear evidences have been shown that either on phonetic system acquisition or prosodic system acquisition, the L1 of the speaker has significant impact on the formation of foreign accent in L2 speech learning literature. However, these studies only show the L1's effect exists by investigating some specific phonological properties of accented speech, without investigating its relationship to the degree of foreign accent. Furthermore, there is no way to quantize the amount of L1's effect during L2 speech learning. Literatures on the relationship between acoustic measurements and perceived accentedness also ignore the measurements that are related to speakers' L1s. The study in this dissertation proposes a computational model that extracts both segmental and suprasegmental acoustic measurements from accented speech signal to analyze the L1's effect on the

formation of foreign accent. It is expected that how close the selected acoustic measurements in accented speech to L1's patterns are also correlated with the degree of foreign accent, and it is natural to predict the correlation is negative.

2. Literatures on automatic accentedness evaluation only includes acoustic features derived with L2 phonological patterns to predict degree of foreign accent. Considering L1's influence to the formation of foreign accent, this study proposes to add acoustic features that represent the phonological distance between accented speech and L1s to automatic accentedness evaluation systems. It is expected that the performance of automatic accentedness evaluation will be boosted by integrating L1's information compared to only considering the acoustic deviation from L2 phonological patterns.

3. Although some studies have investigated the relative contribution of segmental and suprasegmental measurements to the perception of foreign accents, all of these studies evaluated the transplantation methodology on only one foreign language (Italian, German, French or Spanish), thus ignoring if the relative importance of segmental and suprasegmental measurements depends on the s-peaker's L1s. This study utilizes a computational model to investigate whether segmental or suprasegmental acoustic measurements can better explain the variation of accentedness, and to use the accentedness predictability of segmental and suprasegmental acoustic measurements as indication of the relative contribution to the perception of foreign accent. This study also expects that the relative contribution depends on the contrastive information between speaker's L1 and L2 in segmental and suprasegmental feature space.

Chapter 3

METHODOLOGY OVERVIEW

3.1 Introduction

To answer the research questions and test the hypotheses proposed in chapter 1, this dissertation adopts a computational model to investigate the relationship between acoustic representation of accented speech and perceived accentedness. Given accented speech dataset, the core modules of a computational model include feature extraction algorithms to convert acoustic signal to representations related to perceived accentedness, and data analyses to explore the relationship between acoustic representations and perceived accentedness. A following methodology with 3 steps will be employed:

1. Accented speech recordings collection. The very first step is to acquire accented speech data. Then, the accentedness score of each accented speaker needs to be collected to quantify how strong the foreign accent is for native L2 speakers.

2. Measurements related to perceived accentness will be extracted from the acoustic signals. This involves different feature extraction schemes. Some measurements represent pronunciation characteristics and some represent prosodic characteristics. This study will extract measurements that quantify how close the pronunciation of accented speech is to L2, and measurements that quantify how close the prosody of accented speech is to L1.

3. Statistical data analysis is to examine how the perceived accentedness scores (dependent variables) are decided by those acoustic measurements (independent

variables) extracted from the acoustic signals. It includes correlation analysis between independent variables and dependent variables, and regression analysis between groups of independent variables and dependent variables. This study will do regression analyses with different groups of independent variables, for example group of independent variables that are only related to L2, and group of independent variables that are related to both L1 and L2.

This chapter will mainly focus on the first two parts and following chapters will introduce the details of data analysis and corresponding results. Part of this chapter is excerpted from a conference paper by the author (Tu *et al.*, 2018).

3.2 Data collection

3.2.1 Dataset selection

Many non-native speech datasets have been published in the literature (Raab *et al.*, 2007). However, most of them are either not publicly available nor do not have speakers from several different L1s. To have more control on the datasets, the GMU speech accent archive (SAA) (Weinberger, 2013) was chosen as the data source of the speech recordings used in this dissertation. The SAA provides speech samples recorded by speakers from over 300 different L1s. More than 2000 speakers (there are 600 native English speakers currently) read the same paragraph in English:

Please call Stella. Ask her to bring these things with her from the store: six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.

This paragraph was chosen because it includes all of the phonological features considered part of native English speech (Kunath and Weinberger, 2010). With transcription available, it is also easy to derive fine-grained measurements on small phonological unit with computed start and end time. SAA also provides detailed information of the speaker, including age, gender, birth place, native language, English residence country, length of residence and age of English onset. Part of these information is decisive to their degree of foreign accent. The non-native speech corpus used in this study is a subset of the GMU SAA. Four foreign languages: German (9 females, 21 males), French (15 females, 15 males), Mandarin (15 females, 15 males) and Spanish (15 females, 15 males), each of which has 30 speakers. 30 native English speakers (15 females, 15 males) are also added to the set as control native speakers. The four foreign languages are selected because they have diverse contrastive properties with English in both phonetic and prosodic subspaces. The English residence country is limited to the USA, and native English speakers are also born in the USA. This resulted in 150 speakers in the final dataset. The length of each speaker's recording varies in a range from 15-40 seconds. The sampling rate was reduced to 16kHz from 44.1kHz.

3.2.2 Accentedness score collection

SAA does not provide accentedness scores for their speech recordings. In order to quantify the perceived accentedness score , the best way is to ask native speakers of American English to rate the foreign speakers in the dataset. Considering the time and money cost of on-site data collection, Amazon Mechanical Turk (AMT), which is the most popular online crowdsourcing platform, will be chose to acquire the accentedness scores from multiple native American English speakers. The study by Kunath and Weinberger (2010) also collected accentedness scores for recordings in SAA on AMT, and they reported that the collected ratings were reliable enough.

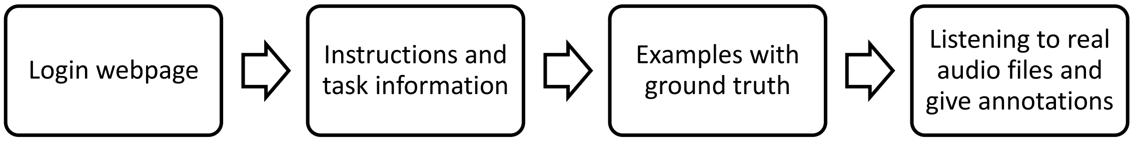

Login webpage
Instructions and task information
Examples with ground truth
Listening to real audio files and give annotations

Figure 3.1: The four steps of the annotation webpage.

The first step of the accentedness score collection is to find annotators to participate in the task, and determine the accentedness score scale. The current accentedness annotation task has several requirements for the annotators: 1) Born in the USA (must be native speaker of American English) 2) Monolingual (only speak American English) 3) Don't speak the four target foreign languages (further make sure they do not speak any of the four foreign languages). 4) No hearing impairment (Make sure they can perceive the foreign accent). 5) At least finished 10 Human Intelligence Task (HIT) [1] that are approved (make sure they have experience using the AMT). 6) HIT approval rate is over 90% on AMT (make sure they devote themselves to each annotation task). Only qualified participants are allowed to do the annotation. To discretize the accentedness, this study employs a four-point scale where one represents no accent/negligible accent, two represents mild accent, three represents strong accent, and four represents very strong accent. This scale has been used in previous collected datasets for example the CSLU: Foreign Accented English datasets (Choueiter *et al.*, 2008), and it is believed that for non-expert annotators a 4-point scale is of less amount of annotation work and higher accuracy compared to a larger scale.

AMT needs an annotation protocol that clearly introduces the whole procedure to finish the annotation task. A website was designed to realize this protocol. The diagram in figure 3.1 shows the whole procedure of the data annotation process.

1. The annotators will first see a webpage (as shown in figure 3.2), asking them to

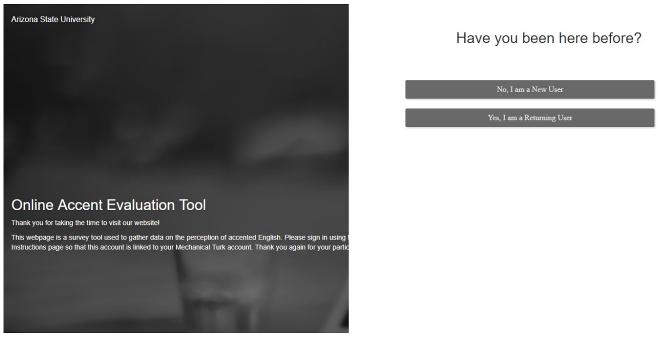
[1] The annotation task on AMT is called HIT.

create a new user or login as a return user. The user ID will be used as identifier to locate their ratings.

2. After finishing step 1, a detailed task instruction and information will be shown to the annotators. The detail is in appendix A. There is also a consent form (in appendix B) for the annotators.

3. Then, four recordings, which are with no accent, mild accent, strong accent and very strong accent respectively, are presented to the annotators for them to get familarization with the 4-point rating scale, as shown in figure 3.3. The groundtruth labels are provided by experienced native American English speakers. This step also enables the annotators be familiar with the content of the recordings.

4. At last, annotators move to the real listening task, as shown in figure 3.4. Each annotator first listens to the recording, then make a choice about the degree of perceived foreign accent and whether the annotator is confident in the response. All 150 speech recordings (including native English speech and accented speech) are randomly permuted in order. If the workers on AMT are asked to listen all of the utterances, the task would take more than 1 hour, and a lot of factors will impact the quality of the collected ratings, such as worker's fatigue (Rzeszotarski *et al.*, 2013). To avoid this, all the utterances are segmented to retain only the first 10 seconds, resulting in 25 minutes listening time for each worker. Previous study (Munro and Derwing, 1995) has shown that the sentence length in the range of 7-13 seconds has little impact on the perceived accentedness. Considering the annotation time and a 2 minutes break, each worker needs to spend about 30-40 minutes for this task. Those annotators finish the task will be rewarded $1.5.


Arizona State University

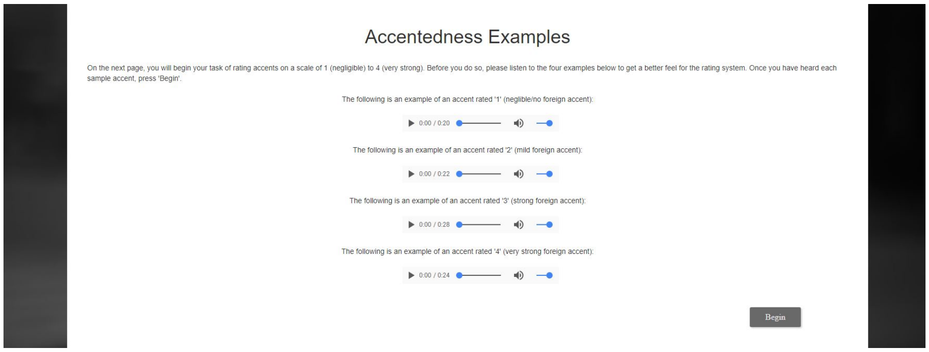
Online Accent Evaluation Tool

Thank you for taking the time to visit our website!

This webpage is a survey tool used to gather data on the perception of accented English. Please sign in using t
Instructions page so that this account is linked to your Mechanical Turk account. Thank you again for your partic

Have you been here before?

Figure 3.2: Annotator's login page.


Accentedness Examples

On the next page, you will begin your task of rating accents on a scale of 1 (negligible) to 4 (very strong). Before you do so, please listen to the four examples below to get a better feel for the rating system. Once you have heard each sample accent, press 'Begin'.

The following is an example of an accent rated '1' (neglible/no foreign accent):

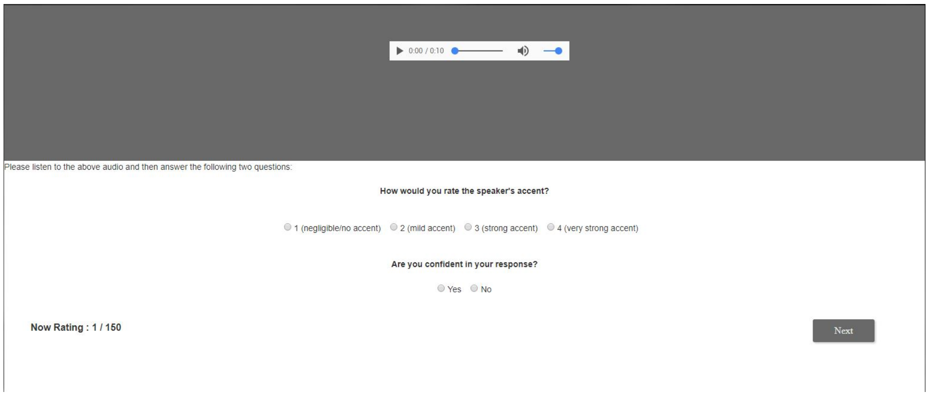
The following is an example of an accent rated '2' (mild foreign accent):

The following is an example of an accent rated '3' (strong foreign accent):

The following is an example of an accent rated '4' (very strong foreign accent):

Figure 3.3: Example accented speech recordings with groundtruth accentedness scores.

Finally, 13 evaluators finished all the listening tasks. The average ratings of all 13 evaluators are taken as the final accentedness rating of each speaker; other studies have used the average of 10 AMT non-expert annotations in other natural language tasks (Snow *et al.*, 2008). The pairwise average inter-rater correlation coefficients are shown in figure 3.5 for each rater, which is calculated by taking the average of the correlation coefficients of the current worker's ratings with other worker's ratings. The average inter-rater correlation coefficients (calculated as the average of all annotators' correlation with other annotators) is 0.73, which is higher enough to prove the consistency of the ratings from 13 evaluators. In figure 3.6, the histograms of the collected ratings across four different foreign languages are presented. It can be found that Mandarin speakers have the strongest accent while German speakers have the mildest accent. This is consistent with expectations considering the phonolog-



Please listen to the above audio and then answer the following two questions:

How would you rate the speaker's accent?

○ 1 (negligible/no accent) ○ 2 (mild accent) ○ 3 (strong accent) ○ 4 (very strong accent)

Are you confident in your response?

○ Yes ○ No

Now Rating : 1 / 150

Figure 3.4: How speech samples are presented to the annotators in listening task.

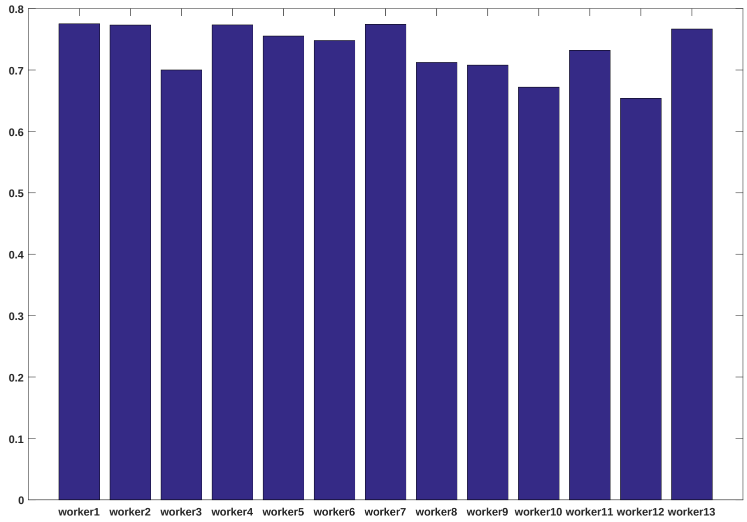


Figure 3.5: Pairwise average correlation coefficients of each worker.

ical similarity between German and English as opposed to other 3 languages. The low accentedness and lack of strongly-accented speakers in the German and French database also means that the variances of the accentedness ratings for these language are relatively low. This poses a challenge in the statistical modeling, which will be further examined in later chapters. In contrast, the average accentedness rating of native English speaker in the dataset is 1.07, which further validate the efectiveness of the annotation.

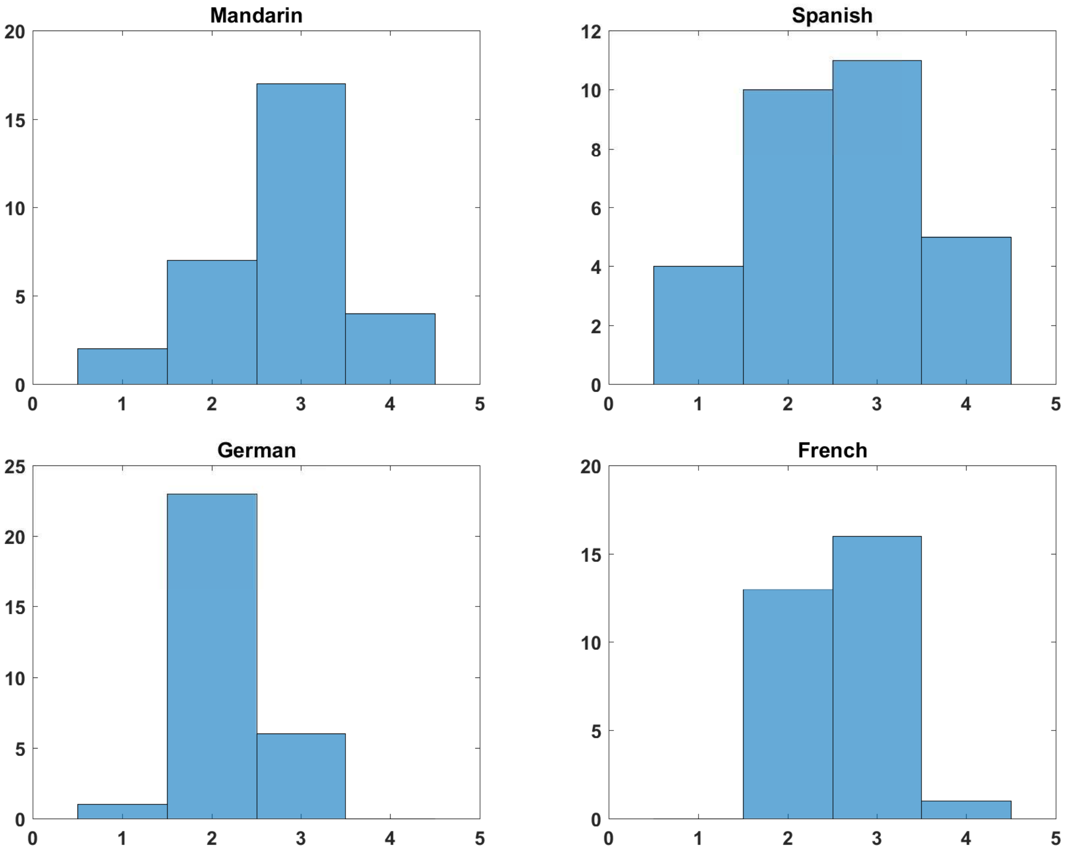


Figure 3.6: Histograms of accentedness scores of different L1s.

3.3 Acoustic analysis

With the accentedness score for each speaker in the accented speech dataset collected, the next step is to extract measurements from the acoustic signal to represent the foreign accent. As mentioned in chapter 1, this study will analyze acoustic measurements in two subspaces: one is characterized by phonetic measurements and the other is characterized by prosodic measurements. Thus, the acoustic analysis here is also done in the two subspaces. Specifically, pronunciation scores of phonemes (including vowels and consonants) and syllables are calculated as representation of the phonetic subspace. The pronunciation scores are computed based on previous studies on phoneme-level goodness of pronunciation (GOP) (Witt and Young, 2000), which

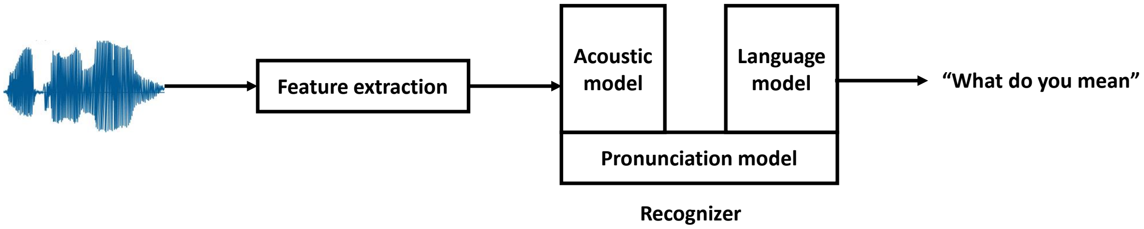


Figure 3.7: Diagram of a typical ASR system. The content of the speech signal is “what do you mean”.

relies on an already-trained automatic speech recognition (ASR) system on native L2 speech. Prosodic measurements are calculated based on the studies by Ramus *et al.* (1999); Grabe and Low (2002). In this dissertation, more prosodic measurements are included as in (Lai *et al.*, 2013). Furthermore, the main contribution of this study is to investigate the relationship between L1 related acoustic measurements and accentedness scores. To calculate L1 related acoustic measurements, corpus of different L1s (German, French, Spanish and Mandarin in this study) are also needed. The remaining part of this section will first briefly review the basic concepts of an ASR system. Then, L1 corpus used in this study will be introduced. Finally, acoustic feature extraction scheme for both phonetic and prosodic information will be presented.

3.3.1 A brief introduction to ASR

Basically, ASR is trying to recognize the content, i.e. what the speaker is saying, in a speech signal. It requires knowledge from different fields, including psychoacoustics, signal processing, linguistics and machine learning [2] . A simplified diagram of an ASR system is shown in figure 3.7. The input waveform is first analyzed within short windows (e.g. 25ms), which are also referred to as frames. Frame based analysis of speech signal is based on the assumption that spectral information is stationary in a

[2] Recent developments on ASR mainly focus on the machine learning part.

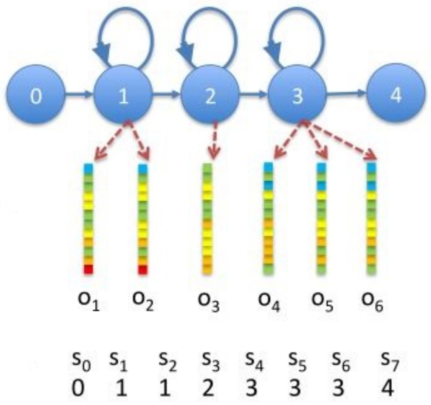

0
1
2
3
4
o_1 o_2 o_3 o_4 o_5 o_6
s_0 s_1 s_2 s_3 s_4 s_5 s_6 s_7
0 1 1 2 3 3 3 4

Figure 3.8: How an HMM aligns input feature vectors with the output state sequence.

short window. This process is done frame by frame. Then, a feature (usually based on Discrete Fourier Transformation, DFT) vector is calculated to represent the spectral information in each frame. The 1-dimensional time domain signal is then converted to a 2-dimensional time-frequency representation (DFT dimension $\times$ number of frames). Since phonemes can be discriminated based on spectral information in acoustic signal, this feature vector is believed to carry information of the identity of phonemes.

Then, the feature vectors of a sentence are sent to a recognizer, which includes three parts: acoustic model, language model and pronunciation model. The acoustic model builds the relationship between feature representation and phonemes. A sequential machine learning model called Hidden Markov Model (HMM) is employed to learn the dynamic transition from one phoneme to another based on the observed feature vectors by aligning each frame with a state of an HMM model (Rabiner, 1989). The reason to use HMM is that the number of frames is different from the number of phonemes in a sentence. There must be a way to correspond each frame to a sub-phoneme unit, which is called a state in an HMM. Each HMM models one phoneme, and the final acoustic model will have many HMMs. As the HMM shown in figure 3.8, input feature vectors o_1 to o_6 are mapped to a state sequence s_1 to s_6, each of which corresponds to a state ID in $\{1, 2, 3\}$. At each frame, it either moves to the next state or stay at the current state. State 0 and 4 are the entrance and exit states of the HMM, which allows transition from a previous phoneme and exit from the current phoneme. Most of the time, a triphone(for example [k-ae+t], [k] is to the left of [ae] and [t] is to the right) instead of a single phone([ae]) is used as the basic modeling unit of an HMM, for the reason that triphone can better model the coarticulation among neighboring phonemes and improve model capability. The possibility a feature vector in one frame is generated by a specific HMM state is modeled with a Gaussian Mixture Model (GMM) or Deep Neural Network (DNN). The possibility a state is transited from another state is decided by an HMM. In summarization, the acoustic model converts a sequence of feature vectors to a sequence of HMM states (with GMM-HMM or GMM-DNN models), and then to phoneme sequence according to the mapping from HMM states to phonemes.

The language model is to convert phoneme sequences output by acoustic model to feasible word sequences, which complies with human usage of words. The pronunciation model involves in this process: it gives the phoneme sequence of each single word in a language. In a nutshell, the pronunciation model is just the lexicon(or dictionary) of a language in most of the time. Pronunciation model will also be used to convert word sequences in transcriptions to phoneme sequences during training stage of the acoustic mode. There is another term commonly seen in ASR field: forced-alignment. It refers to the process to find the start and end time of a phoneme, word or even sentence in a speech signal given the transcription. This can be achieved using acoustic model and pronunciation model. A lot of studies in computational linguistics use forced-alignment to avoid locating phonemes and words in a speech signal by hand. Practically, Kaldi toolkit (Povey *et al.*, 2011) is the most commonly used software to build an ASR system, and it has been well accepted by both academia and industry.

3.3.2 Native speech corpus

In this study, both the L2 and L1s acoustic models and pronunciation models are needed to extract pronunciation score based phonetic features. To build the L2 acoustic model (English for this study), the LibriSpeech corpus (Panayotov *et al.*, 2015) with 960 hours of native English speech recordings was used, and the corresponding training scripts [3] in the Kaldi toolkit. The final acoustic model is a triphone model trained with GMM-HMM on 960 hours of speech data. The feature input is a 39-dimensional second order Mel-Frequency Cepstral Coefficient (MFCC) with utterance-level cepstral mean variance normalization and linear discriminant analysis transformation.

For Mandarin, the publicly accessible AIShell Mandarin Speech corpus (approximately 150 hours training data) (Bu *et al.*, 2017) and the corresponding Kaldi scripts [4] are used. A pronunciation dictionary is included in the dataset. For the remaining three languages (Spanish, French and German), there are no well organized publicly available data. This study uses data from the Voxforge project, and downloads the speech corpora for French (≈ 30 hours), German (≈ 50 hours) and Spanish (≈ 50 hours). Kaldi scripts [5] for the Voxforge English dataset are adapted to train the acoustic models of the three foreign languages. The dictionary for these three languages are from the CMU Sphinx system (Download available [6]). Compared to English and Mandarin acoustic models, the quality of the German, French and Spanish acoustic models trained on Voxforge dataset are not that good (due to varying vocabulary sizes, different number of speakers across languages; some recordings are

[3] https://github.com/kaldi-asr/kaldi/tree/master/egs/librispeech/s5

[4] https://github.com/kaldi-asr/kaldi/tree/master/egs/aishell/s5

[5] https://github.com/kaldi-asr/kaldi/tree/master/egs/voxforge/s5

[6] https://sourceforge.net/projects/cmusphinx/files/Acoustic%20
and%20Language%20Models/

with background noise; pronunciation model is not designed for the datasets). Feature types and structures of acoustic models for the four languages are the same as those used in the English acoustic model.

3.3.3 Pronunciation score based phonetic feature extraction

Features based on the L2 acoustic model

The trained L2 acoustic model can be regarded as the phonetic patterns of native L2 speakers, and it is natural to measure how good non-native L2 speakers' pronunciation is with the native L2 phonetic patterns. Motivated by the work by Witt and Young (2000), the goodness of pronunciation for each phoneme is calculated in the accented speech. To do this, the accented speech is first force-aligned at the phoneme-level using the L2 acoustic model to provide the start and end frame indices of each phoneme. The pronunciation score (PS_{L2}) of the target phoneme p after alignment is defined as

$$\begin{aligned} PS_{\mathrm{L2}}(p) &= \log(P(p|\mathbf{O}^p))/\left|\mathbf{O}^p\right| \\ &= \log\left[\frac{P(\mathbf{O}^p|p)P(p)}{\sum_{q\in Q} P(\mathbf{O}^q|q)P(q)}\right]/\left|\mathbf{O}^p\right|, \end{aligned} \tag{3.1}$$

where $\mathbf{O}^p$ is the feature matrix of phoneme p, $|\mathbf{O}^p|$ is the number of frames of phoneme p after alignment, and Q is the set of all phonemes. If we assume equal priors for all phonemes, we approximate the denominator in Eq. 3.1 with max operator,

$$PS_{\mathrm{L2}}(p) = \log\left[\frac{P(\mathbf{O}^p|p)}{\max_{q\in Q} P(\mathbf{O}^q|q)}\right]/\left|\mathbf{O}^p\right|. \tag{3.2}$$

The conditional likelihood of each phoneme (given the speech frames of the corresponding aligned segment) can be calculated by decoding the sequence of speech features using the L2 acoustic model. It is clear that if the most likely phoneme returned by the acoustic model is the same as the target phoneme p, then $PS_{\mathrm{L2}}(p) = 0$;

otherwise, this value will be negative. The interpretation is that the closer $PS_{\text{L2}}(p)$ is to zero, the closer the pronunciation of phoneme p is to that of native speakers.

L1 acoustic model based measurements

Similarly, the trained L1s acoustic models can be regarded as the phonetic patterns of the L1s of accented speakers. These phonetic patterns can decide how much the pronunciation of L2 is influenced by accented speakers' L1s. In contrast to the PS_{L2} score, there is no transcript to measure the pronunciation of the phonemes in L1. We define a new way to calculate the pronunciation score with the L1 acoustic model which quantifies how close the pronunciation of a phoneme in L2 is to a specific phoneme in L1. The forced-alignment calculated with the L2 acoustic model is used here. The speech frames are first decoded with the L1 acoustic model and find the state path with the highest likelihood. In the path, the corresponding phonemes of each HMM state are recorded and the phoneme with the highest occurrence is considered as the most likely L1 phoneme for a given speech segment. Then, the pronunciation score is calculated as

$$PS_{\text{L1}}(p) = \left[\sum_{t \in T_p} \log \frac{\sum_{s \in S_p} P(o_t|s)}{\sum_{s \in S} P(o_t|s)} \right] / \left| T_p \right| , \tag{3.3}$$

where o_t is the feature vector for frame t and p is the phoneme with the highest occurrences in the best decoding path of the current segment. T_p is the set of frames where each frame corresponds to an HMM state of phoneme p. S_p is the set of HMM states that belong to phoneme p and S is the set of all HMM states. $PS_{\text{L1}}(p)$ essentially quantifies the confidence of the L1 acoustic model that phoneme p was produced for a speech segment. With equation 3.3, a pronunciation score based on the L1 acoustic model can be calculated for each phoneme segment in the original alignment. The implementations of both feature sets are available on Github [7] .

Sentence-level integration

Previous introduced feature extraction methods will output both PS_{L2} and PS_{L1} on phoneme level. However, accented speech of each speaker is a sentence. Thus, a sentence-level integration method is proposed to convert phoneme-level pronunciation scores to a sentence-level feature vector. Specifically, after phoneme-level features $PS_{\text{L2}}(p)$ and $PS_{\text{L1}}(p)$, are extracted, a sentence-level feature extraction scheme was used to convert phoneme-level measurements to a feature vector with a fixed dimension for each utterance. The pronunciation features for vowels, consonants and syllables are first grouped together, and four statistics for each of these three phonetic categories are then calculated: for both $PS_{\text{L2}}(p)$ and $PS_{\text{L1}}(p)$, the minimum, mean, standard deviation and mean-normalized standard deviation (standard deviation divided by mean) of phoneme-level pronunciation scores of vowels, consonants and syllables in each utterance are calculated (implementation available [8]). This results in a total of 12 utterance-level features, and a total of 24 utterance-level features combining both pronunciation information from L1 and L2 acoustic models.

3.3.4 Prosodic feature extraction

To represent speech prosody, durational rhythmic measurements of phonemes and syllables are adopted as the studies by Ramus *et al.* (1999); Grabe and Low (2002). Specifically, an extended speech rhythmic feature set proposed in (Lai *et al.*, 2013) are employed in this study. First, the same forced-alignment results achieved in previous section is reused here to get the start and end time of each phoneme. Then, the

[7] https://github.com/tbright17/kaldi-dnn-ali-gop

[8] https://github.com/tbright17/accent-feat

following measurements are calculated:

1. Mean, standard deviation and mean-normalizd standard deviation (standard deviation divided by mean) of durations of vowels, consonants and syllables.

2. Duration proportion of vowels, consonants and syllables, calculated as the total length of vowels, consonants and syllables divided by the length of the sentence (with starting and ending silence removed).

3. Raw Pairwise Variability Index (rPVI) of durations of vowels, consonants and syllables, calculated as:

   $$rPVI = \sum_{k=1}^{m-1} |d_k - d_{k+1}|/(m-1), \tag{3.4}$$

   where d_k is the duration of kth phoneme or syllable and m is the total number of phonemes or syllables in a sentence.

4. Normalized Pairwise Variability Index (nPVI) of durations of vowels, consonants and syllables, calculated as:

   $$nPVI = \sum_{k=1}^{m-1} |\frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2}|/(m-1), \tag{3.5}$$

   where the notations are the same as in equation 3.4.

Finally, a 18-dimensional feature vector can be extracted from each speech signal. In this study, this rhythmic feature extraction scheme is applied to both L1 speech, L2 speech and accented speech to do contrastive analysis between accented speech and L1, and between L2 speech and accented speech. For the four foreign languages, 1000 sentences with more than 40 phonemes in each are randomly selected from the corresponding native speech corpus. In order to achieve better forced-alignment, another forced-alignment tool (McAuliffe *et al.*, 2017) is employed to align phoneme

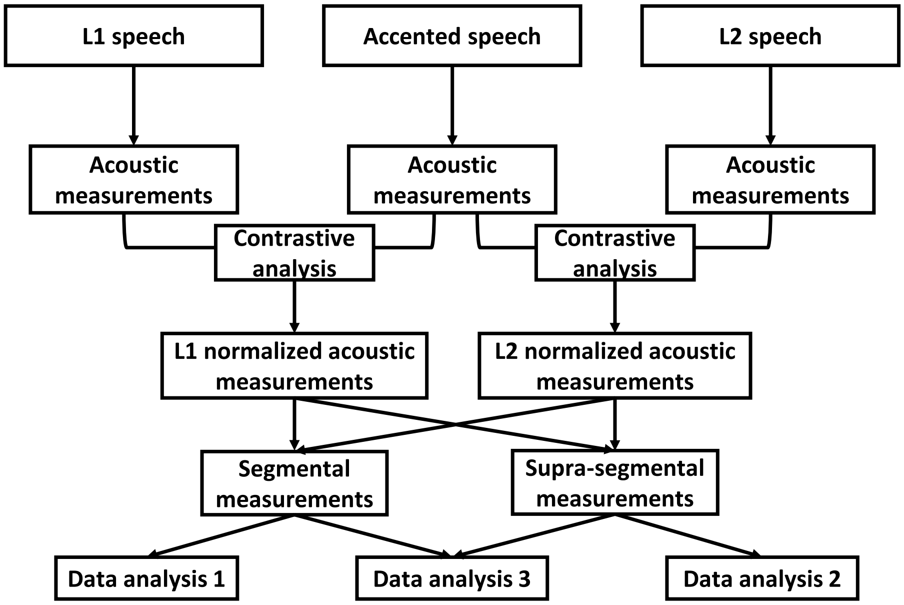


Figure 3.9: Diagram of the methodology used in this study.

sequences with speech recordings because it comes with well-trained foreign language acoustic models (However, because of lack of information, it can not be used to do the computation in section 3.3.3). For native English speech, those measurements are directly calculated on the 30 native American English sentences from SAA using the native English acoustic model trained on Librispeech dataset. The average of rhythmic features of each language (four L1s and English) will be used as the speech prosodic patterns of those languages.

3.4 Procedure

The diagram of the methodology used in this study is shown in figure 3.9. After extracting acoustic measurements from native L1 speech, accented speech and native L2 speech, differential analysis, the goal of which is to quantify the difference between two sets of features, is applied to the L1-accented pair and L2-accented pair. Then, two sets of features can be obtained: the L2 normalized acoustic measurements represent how close the phonological properties in accented speech is to native L2 speech;

the L1 normalized acoustic measurements represent how close the phonological properties in accented speech is to native L1 speech. The segmental feature extraction scheme in section 3.3.3 directly output the L1 and L2 normalized acoustic measurements. This is because the input to that scheme in this case is accented speech and L1 or L2 phonetic patterns (defined by L1 or L2 acoustic models), and the output can represent the difference between accented speech and L1/L2 phonetic patterns. In contrast, differential analysis needs to be done for the suprasegmental feature extraction method in section 3.3.4. The L2 and L1 normalized feature sets can be further categorized into segmental measurements and suprasegmental measurements. The first data analysis, which will be introduced in chapter 4, will investigate the effect of L1 phonetic patterns on the perception accented speech. The second data analysis, which will be introduced in chapter 5, will investigate the effect of L1 prosodic properties on the perception of accented speech. The third data analysis, which will be introduced in chapter 6, will investigate the effect of L1 phonetic and prosodic patterns on the perception of accented speech, and propose a new computational model to do automatic accentedness evaluation. The data analysis methods used in this study are mainly correlation analysis, which examines how the acoustic measurements and accentedness score are correlated, and multiple regression analysis, which examines how well the combination of multiple acoustic measurements can predict the accentedness score. The whole data analysis procedure involves feature preprocessing, feature selection and mode regularization, which will be introduced in detail in later chapters.

Chapter 4

L1'S EFFECT ON PHONETIC PROPERTIES OF ACCENTED SPEECH

4.1 Introduction

This section will investigate the statistical relationship between the phonetic acoustic measurements extracted from accented American English speech (independent variables) and the perceived accetendenss score provided by native American English speakers (dependent variables). Two sets of features will be used as independent variables: one is the pronunciation score based features extracted only using L2 acoustic model, and the other one is the pronunciation score based features extracted using both L1 and L2 acoustic models. This corresponds to the data analysis 1 in figure 3.9 using only L2 normalized segmental acoustic measurements, and the combination of both L1 and L2 normalized segmental acoustic measurements. First, correlational relationship between independent variables and dependent variables is investigated. Second, multiple regression analysis will be employed to analyze how well each set of features can predict the accentedness scores. Results and discussion are in the final part.

4.2 Methods

For each foreign language, the correlation analysis will be done between each dimension of the feature vector and the accentedness scores (average of all 13 annotators). The correlation analysis is done L1 dependently in hope that some L1 specific information will be revealed for testing the second hypothesis in 1. The Pearson correlation coefficients and the corresponding p-value for testing non-correlation will be

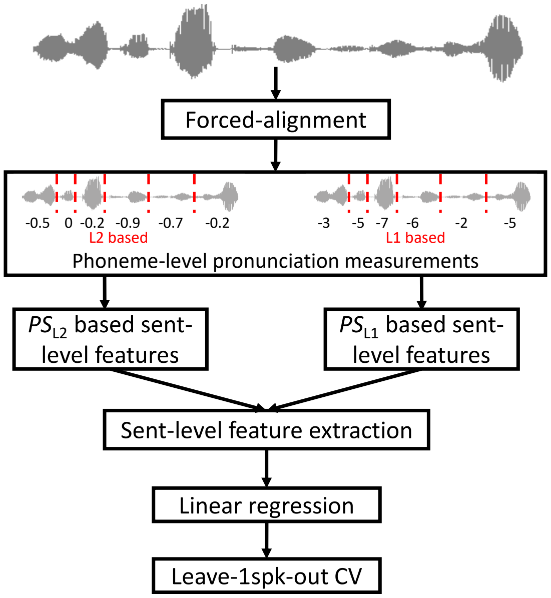

Forced-alignment
-0.5 0 -0.2 -0.9 -0.7 -0.2
L2 based
-3 -5 -7 -6 -2 -5
L1 based
Phoneme-level pronunciation measurements
PS_{L2} based sent-level features
PS_{L1} based sent-level features
Sent-level feature extraction
Linear regression
Leave-1spk-out CV

Figure 4.1: Diagram of the procedure for multiple regression analysis between pronunciation based acoustic measurements and accentedness score.

calculated in this part. Higher correlation coefficients means better correlation, and lower p-value means correlation is more significant.

The whole procedure of multiple regression analysis is shown in figure 4.1. The upper part of the figure shows the feature extraction scheme in section 3.3.3, and will not be described here. Each speaker has a 12-dimensional feature vector quantifying how close the pronunciation is to the L2, and another 12-dimensional feature vector quantifying how close the pronunciation is to the L1. After extracting utterance-level features for all speakers, each speaker has a feature vector and a corresponding accentedness score (in the range of 1 to 4). For speakers that belong to the same L1 category, a linear regression model with a L2-norm regularizer (or ridge regression) is built with data from 29 speakers used to train the model and the remaining speaker used to evaluate the model. The feature vectors are mean and variance normalized first. Feature selection based on univariate linear regression test (Saeys *et al.*, 2007) is also used to select the most predictable features. Basically, the feature selector calculates a score (based on the correlation coefficients between independent variables and dependent variables) for each independent variables given labels in training set and select the independent variables with highest scores. The scikit-learn toolkit is used to implement feature normalization, feature selection and ridge regression (Pedregosa *et al.*, 2011). To generate accentedness predictions for all speakers, leave-one-speaker-out CV based evaluation is performed; this means that a feature selector and a ridge regression model is trained on all combinations of 29 speakers out of 30 speakers, and tested on the 1 remaining. For different input features (12-dimensional utterance-level features or 24-dimensional utterance-level features), the hyperparameters are tuned to achieve the best performance.

As mentioned in section 3.2, the accentedness label distributions for German and French speakers do not span the 1-4 rating scale uniformly. The initial result reveals that the model performance on German and French speakers was comparatively lower (but there was still improvement by adding the feature vector extracted using L1 acoustic model). In an attempt to train our model with more uniformly distributed labels, the German speakers are randomly downsampled from 30 to 18 and French speakers from 30 to 22 in an attempt to uniformly sample the labels. For other two languages, there are still 30 speakers in the results. The Pearson correlation coefficient (PCC, higher better) and the mean absolute error (MAE, lower better) are used to measure the relationship between model prediction and human scores.

4.3 Results

4.3.1 Results of correlation analysis

In table 4.1, PCC (first line) together with p-value (second line) between acoustic measurements extracted from L1 and l2 acoustic models and accentedness scores of four different foreign languages are presented. The results of German and French

Table 4.1: Pearson correlation coefficients (first line) together with p-value (second line) between acoustic measurements extracted from L1 and l2 acoustic models and accentedness scores of four different foreign languages.

	Based on L2 AM				Based on L1 AM			
	German	French	Mandarin	Spanish	German	French	Mandarin	Spanish
Minimum of	-0.36	0.17	-0.53	-0.17	-0.14	0.12	0.12	0.35
vowels' PS	1.48E-01	4.60E-01	2.78E-03	3.58E-01	5.81E-01	6.08E-01	5.27E-01	5.89E-02
Minimum of	-0.45	-0.37	-0.11	-0.47	0.23	0.02	-0.03	0.07
consonants' PS	6.16E-02	1.02E-01	5.69E-01	9.26E-03	3.64E-01	9.28E-02	8.82E-01	7.30E-01
Minimum of	0.00	0.08	-0.33	-0.20	-0.10	0.23	0.31	0.20
syllables' PS	9.94E-01	7.33E-01	7.69E-02	2.82E-01	7.01E-01	3.16E-01	9.80E-02	2.98E-01
Average of	-0.48	-0.27	-0.64	**-0.69**	-0.25	0.40	0.55	**0.53**
vowels' PS	4.31E-02	2.28E-01	1.43E-04	**2.21E-05**	3.14E-01	7.17E-02	1.52E-03	**2.68E-03**
Average of	-0.50	-0.55	-0.64	**-0.69**	0.12	**0.60**	0.33	0.31
consonants' PS	3.60E-02	1.04E-02	1.40E-04	**2.08E-05**	6.26E-01	**3.99E-03**	7.18E-02	9.56E-02
Average of	-0.44	-0.52	-0.68	-0.68	-0.02	0.59	**0.56**	0.44
syllables' PS	6.86E-02	1.59E-02	3.07E-05	2.86E-05	9.29E-01	4.81E-03	**1.17E-03**	1.40E-02
STD of	**0.51**	-0.11	0.61	0.43	**0.35**	-0.24	-0.47	-0.31
vowels' PS	**3.02E-02**	6.41E-01	2.76E-04	1.62E-02	**1.59E-01**	2.85E-01	9.05E-03	9.40E-02
STD of	0.45	0.44	0.40	0.61	-0.20	-0.02	0.38	-0.16
consonants' PS	6.31E-02	4.76E-02	3.05E-02	3.45E-04	4.19E-01	9.38E-01	4.00E-02	3.88E-01
STD of	0.23	0.06	0.43	0.34	0.13	-0.28	-0.10	**-0.53**
syllables' PS	3.49E-01	7.95E-01	1.87E-02	6.48E-02	5.99E-01	2.23E-01	5.94E-01	**2.49E-03**
STD_norm of	0.30	0.63	0.36	0.52	-0.26	-0.01	-0.22	-0.23
vowels' PS	2.34E-01	2.17E-03	5.30E-02	2.97E-03	2.94E-01	9.57E-01	2.35E-01	2.20E-01
STD_norm of	0.35	0.48	**0.73**	0.56	0.24	-0.32	-0.44	-0.01
consonants' PS	1.51E-01	2.77E-02	**4.97E-06**	1.22E-03	3.30E-01	1.60E-1	1.54E-02	9.73E-01
STD_norm of	0.50	**0.69**	0.69	0.57	-0.19	0.06	-0.33	0.21
syllables' PS	3.58E-02	**5.29E-04**	2.82E-05	9.85E-04	4.56E-01	7.81E-01	7.26E-02	2.63E-01

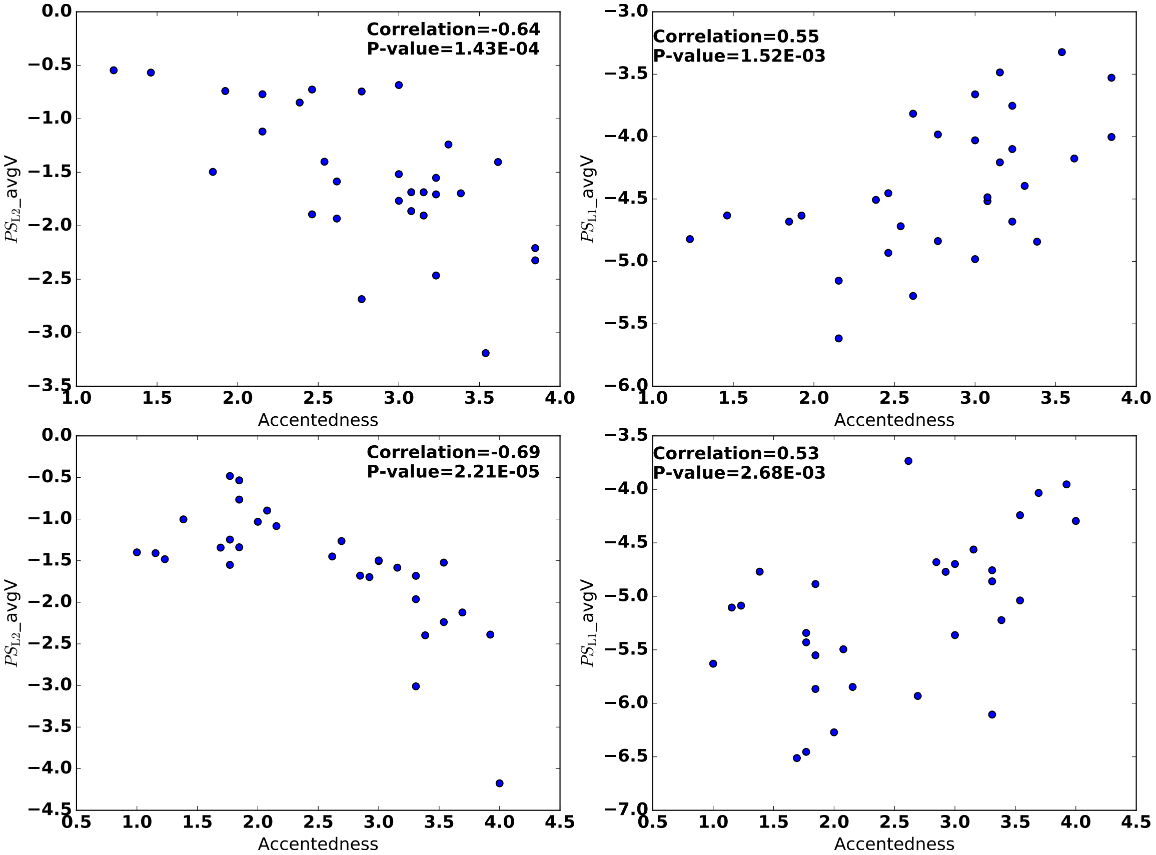


Figure 4.2: Scatter plots between accentedness scores and one dimension of features for Mandarin (first row) and Spanish (second row) speakers.

speakers are based on a downsampled subset because it is found that with the original 30 speakers the correlation coefficients are relatively low. Similar downsampling is also used in multiple regression analysis. In the table, "AM" is short for "acoustic model"; "PS" is short for "pronunciation score"; "STD" is short for "Standard deviation"; "STD_norm" is short for "mean normalized standard deviation". For each feature set (based on L2 AM or based on L1 AM), the highest correlation coefficient between each feature and accentedness scores is in bold for all four foreign languages. From the table, there are several interesting observations:

1. For minimum and average based features (row 3 to row 8), the correlation coefficients with low p-value (means significantly correlated) achieved with L2

acoustic model are negative, while those achieved with L1 acoustic model are positive in most cases. This can be interpreted by the physical meanings of the two feature sets. As introduced in section 3.3.3, for these pronunciation score based features, higher value means closer to pronunciation patterns model by corresponding acoustic models. Thus, for minimum and average features extracted with L1 acoustic model, higher value means the pronunciation of English is closer to pronunciation patterns of the L1; If a speaker is using the pronunciation patterns of his L1 to produce English, very possibly he has a high accentedness score (towards 4 on the scale); Thus, the correlation coefficients between minimum and average features and accentedness scores are positive. On the contrary, for minimum and average features extracted with L2 English acoustic model, higher value means the pronunciation with accent is closer to pronunciation patterns of native English speakers; Thus, the correlation coefficients are negative. In most cases, the correlation coefficients of minimum features are relatively low while the correlation coefficients of average features are relatively high, which tells that the accentedness score can not be determined by one or two phonemes with very low pronunciation scores in a whole utterance.

2. For STD features (row 9 to row 11), the patterns are on the other side compared to minimum and average features. This is also easy to interpret: higher STD means there are some very low pronunciation scores; In terms of features extracted with L2 acoustic models, this means possibly higher accentedness score; In terms of features extracted with L1 acoustic models, this means possible lower accentedness score. The correlation coefficients of STD features are also relatively low compared to average features.

3. STD_norm features (row 12 to row 14) extracted with L1 acoustic model are not very correlative with accentedness score. However, those extracted with L2 acoustic models can have very high correlation coefficients (such as Mandarin speakers). STD_norm features are calculated by dividing values of STD features with values of average features. Ideally, it should have same correlational pattern with STD features considering average features and STD features are oppositely correlated with accentedness score.

4. While the features achieved with L2 acoustic models have higher correlation coefficients with accentedness score, features extracted with L1 acoustic models also show high correlations. This partly supports the first hypothesis in chapter 1, at least in phonetic subspace. In figure 4.2, the scatter plots between average of vowels' PS and accentedness score are presented for Mandarin and Spanish speakers. Based on this observation, it is more likely that when combining features extracted with both L1 and L2 acoustic models can better fit the accentedness score.

4.3.2 Results of multiple regression analysis

In table 4.2, both the PCCs and MAEs between model predicted accentedness and human annotated accentedness for 4 groups of speakers are presented. The results of German and French speakers before down-sampling are also showed in the parentheses. There is a clear improvement when adding L1 acoustic model based features for all 4 L1s. These results show that there is an improvement in model performance consistently and across all languages after adding features from the L1 acoustic model. It proves that the L1 contrastive information between accented speech and L1 can provide extra information for accentedness prediction. This is despite the fact that the annotators know little about the acoustic properties of the speakers'

Table 4.2: PCCs and MAEs between predicted accentedness and human scores for speakers of 4 different L1s.

	PS_{L2} only		PS_{L2} and PS_{L1}	
	PCC	MAE	PCC	MAE
Mandarin	0.707	0.343	**0.727**	**0.329**
Spanish	0.681	0.535	**0.730**	**0.464**
German	0.734 (0.082)	0.204 (0.301)	**0.833** **(0.144)**	**0.163** **(0.287)**
French	0.531 (0.254)	0.335 (0.406)	**0.619** **(0.411)**	**0.303** **(0.370)**

L1s.

In order to show that features extracted with L1 acoustic model really helps with predicting accentedness scores, in table 4.3, L1 acoustic model based features that are selected to predict accentedness scores are showed. Since the multiple regression analyses are done language-independently, different sets of features are selected for different languages, and the number of features selected for each language is also p-resented in the table. Note that for German and French, the feature selection results are based on subsets of speakers after downsampling. It can be found that for all four languages, the average pronunciation score of vowels, consonants and syllables together with minimum of vowels' pronunciation score and standard deviation of vowels' pronunciation scores are selected. This indicates that the first order information of pronunciation scores extracted with L1 acoustic model can help predict the accentedness score. The results of the multiple regression analysis further validate the first hypothesis in chapter 1 that L1-related acoustic measurements can help explain variation in accentedness scores.

Table 4.3: Selected features that are extracted with L1 acoustic model for each language. “num_feature” stands for the total number of selected features by feature selection.

	German (num_feat=24)	French (num_feat=16)	Mandarin (num_feat=14)	Spanish (num_feat=14)
Minimum of vowels’ PS	Yes	Yes	Yes	Yes
Minimum of consonants’ PS	Yes			
Minimum of syllables’ PS	Yes	Yes		
Average of vowels’ PS	Yes	Yes	Yes	Yes
Average of consonants’ PS	Yes	Yes	Yes	Yes
Average of syllables’ PS	Yes	Yes	Yes	Yes
STD of vowels’ PS	Yes	Yes	Yes	Yes
STD of consonants’ PS	Yes			
STD of syllables’ PS	Yes	Yes		Yes
STD_norm of vowels’ PS	Yes			
STD_norm of consonants’ PS	Yes	Yes	Yes	
STD_norm of syllables’ PS	Yes		Yes	

4.4 Discussion

The results in table 4.2 reveal that the improvement in performance of regression model varies across different L1s. There are several possible reasons for this including the different modeling quality of the L1s' ASR systems, the accentedness annotation quality, or the contribution of articulation features to perceived impressions of accentedness for different languages. Another interesting aspect that is worthy of additional investigation is that although there is knowledge transfer from L1 to L2 during L2 acquisition, this influence can vary across different L1s and even different speakers. For example, some research suggests that there exist some universal effects in L2 speech learning process that are independent of a speaker's L1 Chang (2010). The approach in this study may provide a means of comparing L1-specific and L1-agnostic pronunciation errors in an attempt to computationally identify some of the universal effects. Specifically, comparing the L1 and L2 acoustic pronunciation scores of English phonemes produced by L2 learners can indicate which English phonemes are not pronounced well due to the speaker is using a similar way with phonemes in L1 phonetic system (high L1 acoustic model based pronunciation score), and which English phonemes are not pronounced well but they also have low L1 acoustic model based pronunciation score (means the pronunciation pattern has nothing to do with the L1 phonetic system).

It has been shown that the proposed feature sets can boost the performance of accentedness prediction. However, there is still room for improvement. First, as mentioned previously, the GMU speech accent archive dataset has a limited number of speakers and small variation of accentedness for some languages. The recording environment also varies by speaker. A cleaner dataset with uniform accentedness ratings is better suited for our application. Second, the amount and quality of training data for L1 acoustic models can be improved since it is quite limited for some of the languages (Spanish, German and French in this study). More accurate L1 acoustic models may result in an improvement of algorithm performance. Third, it is well known that accentedness is related to both pronunciation and prosodic features. This chapter mainly focuses on pronunciation based features. In the next chapter, the same framework will be extended to speech prosody features.

Chapter 5

L1'S EFFECT ON PROSODIC PROPERTIES OF ACCENTED SPEECH

5.1 Introduction

The previous chapter applies the proposed methodology for accentedness perception to pronunciation based segmental features, and proves that integrating L1 pronunciation information by extracting pronunciation scores of accented English speech with L1 acoustic model can improve the prediction accuracy of accentedness perception. This chapter will focus on applying the same methodology to speech prosodic features to study whether L1 prosodic patterns affect the perception of accentedness. As mentioned in chapter 3, durational rhythmic features will be used as proxy of speech prosody. The methods and analysis of results are almost the same as chapter 4. Details will be introduced in following sections.

5.2 Methods

Chapter 3 describes the procedure to extract durational rhythmic features. The extracted rhythmic features for native L1, accented L2 speech and native L2 are represented with $\mathbf{x_{L1}}$, $\mathbf{X_{acc}}$ and $\mathbf{x_{L2}}$ respectively. Note that $\mathbf{x_{L1}}$ and $\mathbf{x_{L2}}$ are vectors because they are the average of features extracted from multiple speech recordings. These three sets of features are converted to accent related features by taking the absolute difference between $\mathbf{x_{L1}}$ and $\mathbf{X_{acc}}$ and $\mathbf{x_{L2}}$ and $\mathbf{X_{acc}}$. $|\mathbf{x_{L2}} - \mathbf{X_{acc}}|$ represent the difference between the rhythmic patterns of accented speech and target L2 speech, while $|\mathbf{x_{L1}} - \mathbf{X_{acc}}|$ represents the difference between the rhythmic patterns of accented speech and speaker's L1 speech. Here, the subtraction is broadcasted to every row of

$\mathbf{X_{acc}}$ to get the contrastive information for each speaker with accent.

The first analysis is the correlation analysis between speech prosodic features and accentedness scores averaged on 13 annotators. Similarity, the PCC is calculated between every features of the 18-dimensional feature vectors in a language-dependent way. The procedure is different from previous chapter observing that the feature dimension is higher than pronunciation features. Thus, only the top-12 features with highest PCC with accentedness scores are shown for each language together with the p-values (lower p-value stands for more statistically significant correlation).

Same multiple regression analysis is done except that the number of input features is changed to 18. Downsampling is not used since it does not help for German speakers. Thus, for each language, multiple regression analysis is conducted between 18-dimensional speech rhythmic measurements and accentedness scores of 30 speakers. Specifically, $|\mathbf{x_{L2}} - \mathbf{X_{acc}}|$ is used as the baseline model which only takes the difference of speech rhythmic patterns between accented speech and L2 into consideration. Then, $|\mathbf{x_{L1}} - \mathbf{X_{acc}}|$ can be combined into the baseline feature set to model the distance between accented speech and L1 on suprasegmental feature space. Finally, input to the baseline model is a 18-dimensional feature vector, and adding L1-related information result in a 36-dimensional feature vector.

As in chapter 4, the input feature vectors are first normalized with mean and standard deviation on each dimension. Then, a feature selector based on univariate regression test is applied to select the most predictable features. Ridge regression is used to learn the relationship between input features and accentedness score. The same leave-one-speaker-out CV is employed to evaluate the performance on accentedness prediction. Hyperparameters including number of features selected and strength of 2-norm regularization in ridge regression are tuned to achieve the best CV performance. PCC and MAE are reported language dependently. To better illustrate the

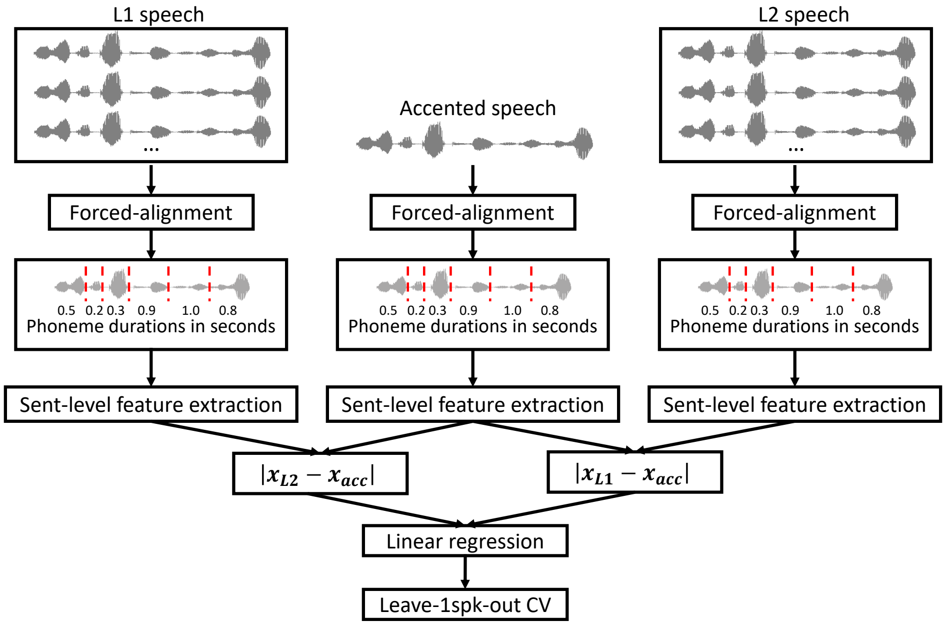


Figure 5.1: Diagram of the procedure for multiple regression analysis between suprasegmental prosodic features and accentedness scores. Here, $\mathbf{x_{L1}}$ and $\mathbf{x_{L2}}$ are the average of features of all speech recordings. $\mathbf{x_{acc}}$ is the feature vector for one accented speech recording.

process, figure 5.1 shows the whole procedure of multiple regression analysis.

5.3 Results

5.3.1 Results of correlation analysis

In figure 5.2, PCCs together with p-values between two sets of speech rhythmic features and accentedness scores of four different foreign languages are presented. German speakers are the from the downsampled subsets with 18 speakers. There is no downsampling for French speakers because the correlation coefficients are not affected by the non-uniform distribution of accentedness scores. Feature names on X-axis are abbreviations: per{V,C,Syl} represents the percentage of durations of vowels, consonants and syllables, avg{V,C,Syl} represents the average durations, std{V,C,Syl} represents the standard deviation of durations, Vacro{V,C,Syl} represents the mean-

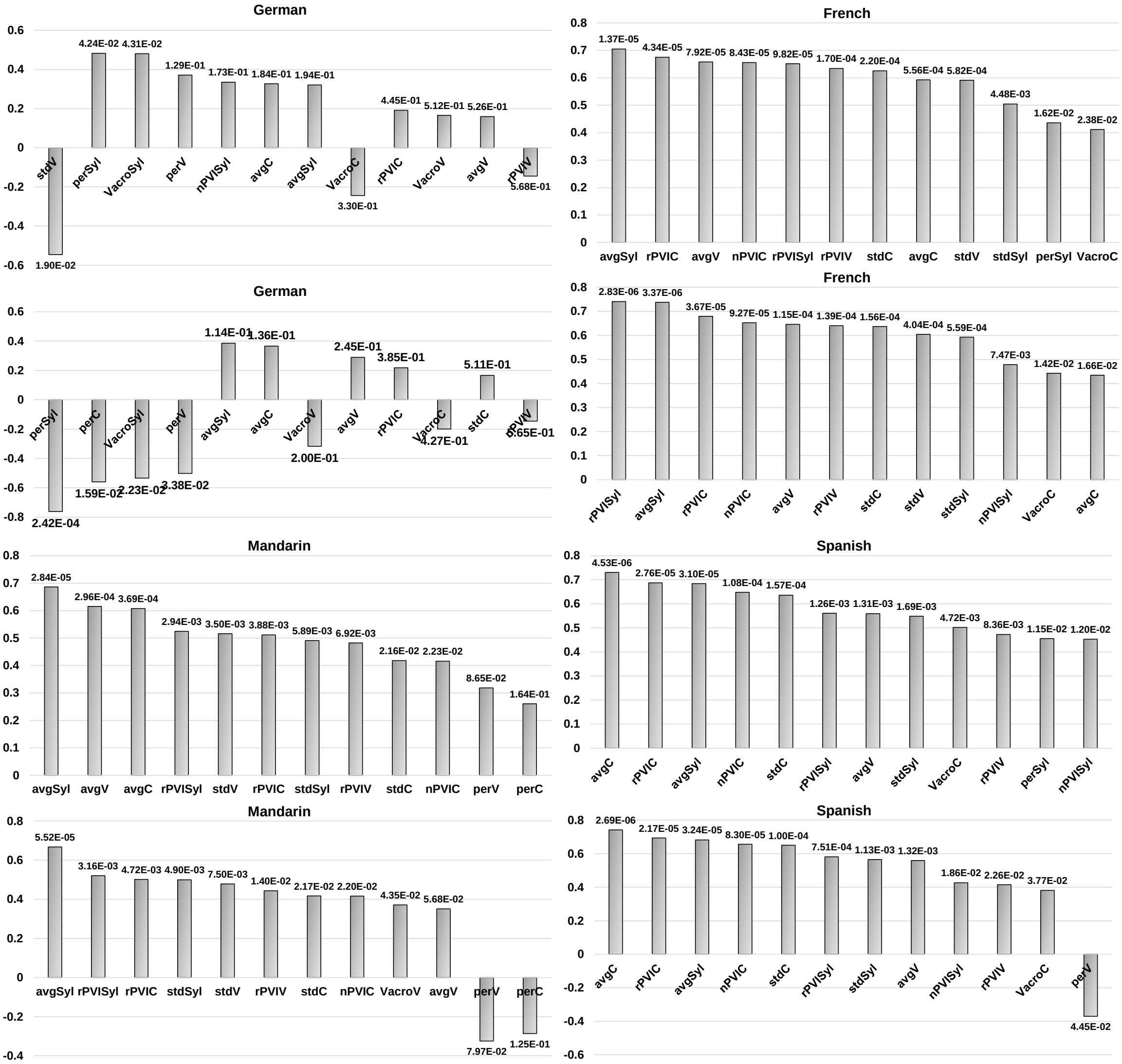


Figure 5.2: Bar plots of the top-12 features highly correlated with accentedness scores in feature sets $|\mathbf{x_{L2}} - \mathbf{X_{acc}}|$ (upper panel in each subfigure) and $|\mathbf{x_{L1}} - \mathbf{X_{acc}}|$ (lower panel in each subfigure). Y-axis includes the correlation coefficients with accentedness score and X-axis includes feature names. The numbers on top of each bar are the p-value for testing non-correlation.

normalizd standard deviation of durations, rPVI{V,C,Syl} represents the Raw PVI of durations and nPVI{V,C,Syl} represents the Normalized PVI) of durations. Several interesting observations can be summarized:

1. Except for German speakers, rhythmic features of other three languages all have relatively high correlation coefficients (>0.6) with accentedness scores. This can be attributed to the similarity of rhythmic patterns between English and German (as shown in figure 2.4 and table 2.4, also in the study by Li and Post (2014)). It becomes hard to use rhythmic features to differentiate between mild and strong accent when the rhythmic patterns of L1 is already very close to L2.

2. As shown in previous chapter, the most predictable features extracted with L1 acoustic models have opposite correlation with accentedness scores compared to features extracted with L2 acoustic models. However, for rhythmic features, it can be found that most features of both $|\mathbf{x_{L2}} - \mathbf{X_{acc}}|$ and $|\mathbf{x_{L1}} - \mathbf{X_{acc}}|$ are positively correlated with accentedness scores (except for German speakers). Only a few dimensions of $|\mathbf{x_{L1}} - \mathbf{X_{acc}}|$ have negative correlation with accentedness scores. This indicates that for some speakers, values of feature dimensions in $\mathbf{X_{acc}}$ are not within the range from values of $\mathbf{x_{L2}}$ to values of $\mathbf{x_{L2}}$ in corresponding dimensions, while values on some feature dimensions are between the values in L1 and L2. This is also observed in the study by White and Mattys (2007). This observation is consistent with the founds in (Li and Post, 2014) where the authors believe that for speech rhythm acquisition there is a multi-systemic model of L2 rhythm acquisition and both transferred L1 knowledge and universal effects independent of L1 played a role.

3. For languages that have high correlation coefficients, it can be found that the average durations features and PVI features are the most correlated ones. This

is also consistent with studies by Ordin and Polyanskaya (2015) where they show the different rhythmic feature values in different proficiency levels: beginners, intermediate and advanced, in spite that they did not provide correlation coefficients between rhythmic feature values and how strong the accent is.

5.3.2 Results of multiple regression analysis

During the experiment, it was found that For French and Mandarin, using feature set $[|\mathbf{x_{L2}} - \mathbf{X_{acc}}|, |\mathbf{x_{L2}} - \mathbf{X_{acc}}|\text{-}|\mathbf{x_{L1}} - \mathbf{X_{acc}}|]$ as the way to integrate L1 information gave the best performance of leave-one-speaker-out CV; for Spanish and German, feature set $[|\mathbf{x_{L2}} - \mathbf{X_{acc}}|, |\mathbf{x_{L1}} - \mathbf{X_{acc}}|]$ gave the best performance. Since for French and Mandarin, using $[|\mathbf{x_{L2}} - \mathbf{X_{acc}}|, |\mathbf{x_{L1}} - \mathbf{X_{acc}}|]$ can also achieve better performance than the baseline model, the difference of the best feature sets across languages is probably due to different speech prosody patterns. In table 5.1, both the PCCs and MAEs between model predicted accentedness and human annotated accentedness for 4 groups of speakers are presented. The results for German speakers are based on the 18 speakers after downsampling (same as chapter 4). However, there is no downsampling for French speakers, because without downsampling, the performance on French speakers is already satisfied. There is consistent improvement when adding L1 rhythmic patterns based features for all 4 L1s. These results show that there is benefit to model performance consistently and across all four languages after adding features from contrastive information with L1 rhythmic patterns. It proves that the rhythmic contrastive information between accented speech and L1 can provide extra information for accentedness prediction. This is also despite the fact that the annotators know little about the acoustic properties of the speakers' L1s.

In order to show that features extracted with L1 rhythmic patterns really helps with predicting accentedness scores, in table 5.2 L1 rhythmic patterns based features

Table 5.1: PCCs and MAEs between predicted accentedness and human scores for speakers of three different L1s.

	$[\lvert \mathbf{x_{L2}} - \mathbf{X_{acc}} \rvert]$		With $\lvert \mathbf{x_{L1}} - \mathbf{X_{acc}} \rvert$	
	PCC	MAE	PCC	MAE
German	0.583	0.202	0.772	0.180
French	0.647	0.310	0.680	0.289
Mandarin	0.581	0.425	0.712	0.380
Spanish	0.698	0.507	0.729	0.482

Table 5.2: Selected L1-related feature dimensions from $[\lvert \mathbf{x_{L2}} - \mathbf{X_{acc}} \rvert$, $\lvert \mathbf{x_{L2}} - \mathbf{X_{acc}} \rvert$-$\lvert \mathbf{x_{L1}} - \mathbf{X_{acc}} \rvert]$ for French and Mandarin speakers or $[\lvert \mathbf{x_{L2}} - \mathbf{X_{acc}} \rvert$, $\lvert \mathbf{x_{L1}} - \mathbf{X_{acc}} \rvert]$ for Spanish speakers. "num_feature" stands for the total number of selected features by feature selection.

	Selected L1-related features
German (num_feat=18)	avgV,avgC,avgSyl,stdC,VacroC,VacroSyl,perV perC,perSyl,rPVIC,nPVIV,nPVIC
French (num_feat=25)	avgC,avgSyl,stdV,stdC,stdSyl,VacroC,perV,perC perSyl,rPVIV,rPVIC,rPVISyl,nPVISyl
Mandarin (num_feat=15)	avgV,avgC,stdV,VacroSyl,perV,rPVIV,rPVIC
Spanish (num_feat=11)	avgV,avgC,avgSyl,stdC,stdSyl,rPVIC,rPVISyl,nPVIC

that are selected to predict accentedness scores are showed. Since the multiple regression analyses are done language-dependently, different sets of features are selected for different languages, and the number of features selected for each language is also presented in the table. It can be found that for French and Spanish speakers, the durational measurements of L1 consonants and syllables are more often selected, while for Mandarin speakers the durational measurements of L1 vowels are more import features. The study by Li and Post (2014) compared durational measurements of Mandarin accented English and native English. They showed that vocalic rhythmic measurements can well discriminate Mandarin learners at different proficiency levels. For French and Spanish speakers, there are no studies showing the progressive change of consonantal and syllable rhythmic measurements along proficiency levels. The results are reasonable considering both French and Spanish are syllable-timed languages while English is stress-timed languages. For German speakers, L1-related vocalic, consonantal and syllabic measurements are all important for accentedness prediction. The results of the multiple regression analysis further validate the first hypothesis in chapter 1.

5.4 Discussion

This chapter shows that the speech prosodic properties transferred from L1 can also help deciding how strong the foreign accent of L2 learners is . This conforms with previous studies, where the authors show the L1's effect on L2 prosody acquisition (Rasier and Hiligsmann, 2007; Stockmal *et al.*, 2005; White and Mattys, 2007; Li and Post, 2014; Ordin and Polyanskaya, 2015). However, based on the correlation analysis in figure 5.2, on most feature dimensions, it does not indicate that if the rhythmic property on that dimension is further from L1, the foreign accent is milder. This is in contrast to the results in table 4.1. The first possible reason is that while previous studies show the effect of L1 on L2 rhythmic pattern acquisition, there are also obvious universal effect that are independent of L1. For example, the study Ordin and Polyanskaya (2015) showed that the PVI measurements of English speech produced by French speakers can be even higher than native English speakers given that English speech has much higher PVI measurements than French speech. The second possible reason is that all the rhythmic measurements in this study are based on automatic forced alignment. For speakers with not very strong accent, there will be much fewer forced-alignment errors. However, for speakers with very strong accent, the forced-alignment results may not be very accurate. This will also affect the correlation analysis between features and accentedness scores.

There are also some interesting implications combining the findings in this chapter and chapter 4. Compared to the results in chapter 4, for German and Mandarin speakers, using only segmental pronunciation based features can better predict accentedness scores than using only suprasegmental rhythmic features; while for French and Spanish, the suprasegmental rhythmic features perform better. Based on the language differential analysis presented in chapter 2, where it shows the relative distances of different L1s to English on both phonetic space and rhythmic space, the results suggest the relative contribution of segmental and suprasegmental acoustic characteristics to the perception of foreign accent for different L1s. For example, German and English have very similar rhythmic patterns, thus segmental measurements characterizing phoneme pronunciation can better discriminate speakers with strong and weak foreign accent. Mandarin is far to English on both phonetic and prosodic subspaces, and the results indicates that for native English speakers, the pronunciation of English phonemes is more decisive to determine the accentedness. Spanish and French are all syllable-timed languages, thus native English speakers attribute the foreign accent of Spanish and French speakers more to suprasegmental inaccuracy than to segmental inaccuracy. In contrast to previous studies using phonological properties transplantation to investigate the relative importance of segmental and suprasegmental in accentedness perception, this study provide a new way to look at the same problem with the advantage that this method can provide quantitative analysis. This study further demonstrates that the relative importance of segmental and suprasegmental features to the perception of foreign accent may vary according to accented speaker's L1 background, and the variation is due to the contrastive patterns between L1 and L2 in segmental and suprasegmental feature spaces. These findings support the second hypothesis in 1, which claims that "phonological properties in different subspaces (phonetic or prosodic) of accented speech produced by speakers from different L1 backgrounds will have distinct contribution to perceived accentedness".

This study shows that with extra speaker's L1 information, the perception of accentedness can be better modeled compared to only using the deviation from native L2. At first thought, this is against intuition, especially considering the annotators in this study do not know the identification of the speakers's L1, neither can speak those L1s. However, previous study (Yuan *et al.*, 2010) has shown that the accentedness perception of non-native speech by non-native L2 speakers has preference over L1 backgrounds. In the study, eight Mandarin judges who were considered as experienced English speakers were asked to rate the accentedness of speakers speaking eight different L1s. The results showed that Mandarin judges tended to underestimate the accentedness of Cantonese and Mandarin speakers the most, followed by German, Japanese and Vietnamese speakers, and French, Spanish and Russian speakers the least. The authors suggested that structural similarities or differences between the L1 languages of the speakers and the listeners play an important role in the listeners' perception of accentedness. Flege *et al.* (1995) reviewed the factors that affecting the perception of accentedness by native L2 speakers. However, since degree of foreign accent is originally defined as the native L2 speaker's perception of deviations from a pronunciation norm that a listener attributes to the talker not speaking the target language natively (McCullough, 2013a), there is no way to study if native L2 speakers have preference depending on the phonological differences and similarities between L1 and L2. This study tends to believe that the difference or similarities on specific phonological dimensions between L1 and L2 play a decisive role in accentedness perception. This effect could vary across speakers (as concluded by Major (1987b) that the amount of L1's influence decreased as learners become more proficient in L2, and this behavior may vary for different learners) instead of simply decided by the distance between averaged L1 phonological patterns and averaged L2 phonological patterns. Chapter 7 will take Mandarin speakers as an example to show the effect of L1 on both phonetic and prosodic properties of accented speech.

Chapter 6

A COMPUTATIONAL MODEL FOR ACCENTEDNESS PERCEPTION WITH L1 INFORMATION

6.1 Introduction

This chapter will use the knowledge derived from previous chapters, and propose a new scheme for automatic accentedness evaluation system. The system features a novel acoustic feature extraction process, which not only combines both segmental and suprasegmental information but also integrates speakers' L1 information. The way of adding L1 information in this study is also novel in automatic accentedness evaluation literature except for one study. Moustroufas and Digalakis (2007) used utterance-level pronunciation scores extracted from both L1 and L2 acoustic models, calculated frame-wise and averaged over the utterance. However, the proposed system has an important difference: the pronunciation scores are calculated on phoneme segments and provide more specific information regarding the accentedness of different phonemic categories. Suprasegmental acoustic measurements representing speech prosody are also included. Furthermore, Moustroufas and Digalakis (2007) assume the human evaluator can speak both L1 and L2 and experiments were conducted on only one L1. This study wants to investigate if the L1 acoustic model can help improve prediction even if the human evaluators have no knowledge of the underlying L1.

To validate the proposed system, the dataset introduced in chapter 3 will be employed again to conduct experiments on automatic accentedness evaluation both L1-dependently and L1-independently. Same leave-one-speaker-out CV will be used

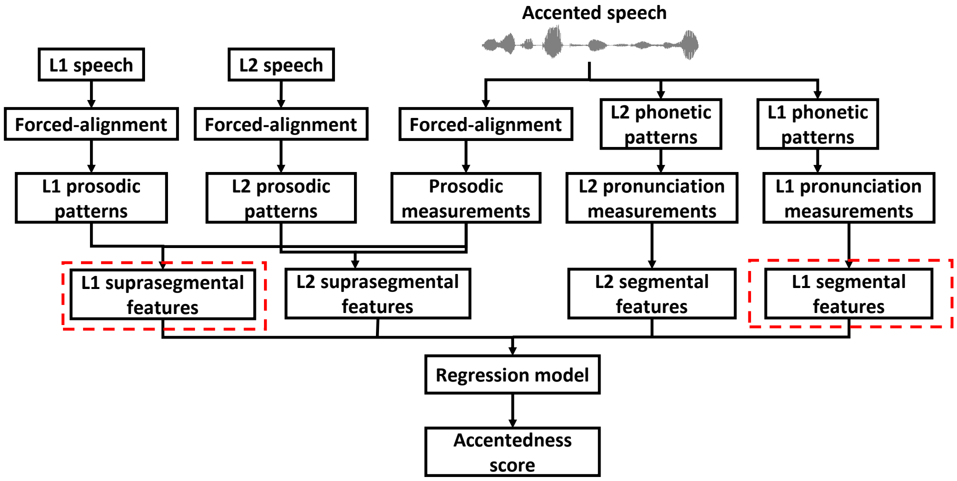


Figure 6.1: Diagram of the proposed computational model. The blocks within red box are the highlights of the current model.

to evaluate the performance of the system. Finally, possible extensions and improvements over the state-of-the-art systems are discussed.

6.2 Method

The diagram of the proposed computational model is shown in figure 6.1. Prerequisites include a well trained acoustic model (hybrid system built on GMM-HMM or DNN-HMM) on native L2 speech, a well trained acoustic model on native L1 speech, a corpus of native L2 speech for extracting L2 prosodic patterns and a corpus of native L1 speech for extracting L1 prosodic patterns. First, accented speech in L2, native L1 speech and native L2 speech are processed with forced-alignment tools to obtain the durations of each phoneme in the transcripts. There are many available forced-alignment tools with open access [1] . Some of them support forced-alignment for multiple languages. For accented speech, usually the forced-alignment performance is inferior compared to native speech because those tools also use acoustic models trained on native speech. To relieve this problem, some recent studies trained the

[1] As summarized in `https://github.com/pettarin/forced-alignment-tools`

acoustic models for accented speech forced-alignment directly on accented speech to achieve data matching (Tao *et al.*, 2016; Qian *et al.*, 2017). However, it requires huge amount of non-native speech recordings which is usually inaccessible. This study employs a forced-alignment tool with an acoustic model trained on a native English speech corpus with about 1000 hours training data to get the phonemes durations of accented speech. After obtaining the phoneme durations of native L1 speech, native L2 speech and accented speech, the feature extraction procedure introduced in section 3.3.4 will be applied to get the prosodic patterns of native L1 and L2. This can be achieved by averaging over the sentence-level features over all native L1 and L2 utterances. At the same time, prosodic feature vectors of each accented speech utterances are saved for following process. What previous studies have investigated is that computing the difference between prosodic measurements of accented speech and native L2 speech gives the deviation from native prosodic patterns. This study improves this by adding the difference between prosodic measurements of accented speech and native L1 speech to represent how much the prosodic patterns of accented speaker are affected by L1. This results in two sets of suprasegmental feature vectors for each accented speech utterance.

In parallel, accented speech recordings are also sent to both L1 and L2 acoustic models to get the pronunciation scores of each phoneme in the utterance based on the corresponding acoustic model. The algorithms proposed in section 3.3.3 are used to convert phoneme-level pronunciation score to sentence level pronunciation measurements. Different from previous studies in the literature, this study not only measures the pronunciation mismatch with native L2 speech but also how much the pronunciation in L2 is affected by the speaker's L1. This again results in two sets of segmental feature vectors for each accented speech utterance. Up to now, there are four feature sets for each accented speech utterance. All of them can be concatenated to form a larger feature vector. Feature vectors without L1 information will be used as baseline system in this study, which has been adopted by recent studies (Black *et al.*, 2015; Tao *et al.*, 2016; Qian *et al.*, 2017).

Almost all studies in literature treat the accentedness (or nativeness) evaluation as a regression problem. With the developed sentence-level features, each speaker becomes a data sample with a labeled accentedness score. The accentedness score can be on different scales depending on tasks. This study use a 4-point scale to annotate the accentedness score. Usually, the label will be the average of multiple annotators to reduce inter-rater variability. With feature representation and labels, a regression model can be trained to learn the mapping from input feature to accentedness score. Considering the relatively small number of speakers, this study adopts ridge regression (linear regression with L2-norm regularization) together with a simple feature selection algorithm based on univariate regression analysis. Depending on the dataset, different regression models can be used to achieve better performance. For example, support vector regression (Black *et al.*, 2015), Gaussian process (Grósz *et al.*, 2015), random forest (Qian *et al.*, 2017) and Deep neural networks (Grósz *et al.*, 2015) are also used in previous studies. SVR is also tried in this study but it is not better than linear regression.

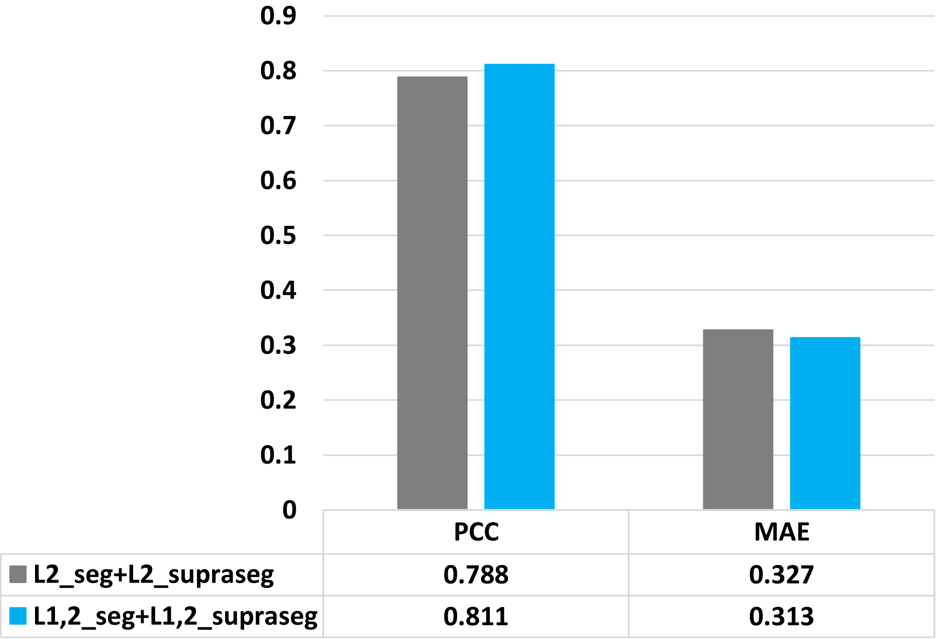
To evaluate the proposed systems, experiments are conducted on both L1-dependent and L1-independent tasks. For L1-dependent task, the system is built on speakers from one L1; for L1-independent task, the system is built on speakers from different L1s. In both cases, leave-one-speaker-out CV is used to evaluate the system's performance because leave-one-out CV is almost the unbiased estimate of generalization error (Elisseeff *et al.*, 2003). As previous chapters, PCC and MAE on all speakers are used as performance indicators.

Table 6.1: Performance of accentedness score prediction with different feature sets for different L1s. "L2_seg" stands for L2 pronunciation features. "+L1_seg" means adding L1 pronunciation features to original L2 pronunciation features. "L2_supraseg" stands for L2 prosodic features; "+L1_supraseg" means adding L1 prosodic features to original L2 prosodic features. "L2_seg+L2_supraseg" represents combining both L2 pronunciation features and L2 prosodic features. "L1,2_seg+L1,2_supraseg" represents combining all four sets of features together.

	PCC				MAE			
	German	French	Mandarin	Spanish	German	French	Mandarin	Spanish
L2_seg	0.734	0.254	0.707	0.681	0.204	0.406	0.343	0.535
+L1_seg	0.833	0.411	0.727	0.730	0.163	0.370	0.329	0.464
L2_supraseg	0.583	0.647	0.581	0.698	0.202	0.310	0.425	0.507
+L1_supraseg	0.772	0.680	0.712	0.729	0.180	0.289	0.380	0.482
L2_seg + L2_supraseg	0.494	0.667	0.733	0.846	0.251	0.308	0.319	0.404
L1,2_seg+L1,2_supraseg	0.590	0.709	0.771	0.898	0.225	0.277	0.296	0.341

6.3 Results

Table 6.1 shows the performance of accentedness score prediction in L1-dependent way with different feature sets. Part of the results is from table 4.2 and 5.1. The results on German speakers are achieved with the 18-speaker subset, and all 30 French speakers are used here. The results show that when combining both segmental and suprasegmental features, the performance is better than either only using segmental features or only using suprasegmental features. For Spanish speakers, the improvement is the largest. However, this observation does not hold for German speakers: the performance degrades a lot after combining both segmental and suprasegmental features. Possible reasons for this could be model overfitting considering there are only 18 data samples, resulting in worse prediction accuracy when combining two feature sets. This problem can be relieved by using an ensemble of two ridge regression model trained on segmental and suprasegmental features separately. Actually, when applying a weighted sum on the predictions of segmental model and suprasegmental



	PCC	MAE
■ L2_seg+L2_supraseg	0.788	0.327
■ L1,2_seg+L1,2_supraseg	0.811	0.313

Figure 6.2: Bar plots and detailed values of results in L1-independent way.

model, the correlation and MAE are 0.791 and 0.171 respectively, which are better than single model. To keep the results consistent, only performance on feature-level fusion are presented. When adding L1 related information to the feature sets, the prediction accuracy is further improved for all four L1s. For Spanish speakers, the PCC is as high as 0.9 (but the MAE is also the highest). Again, the ensemble model performance of German speakers is 0.890 for PCC and 0.137 for MAE, which are also big improvement compared to model trained without L1-related features. This improvement is expected based on the results shown in previous two chapters.

Figure 6.2 shows the results when all speakers from German (18-speaker subset), French, Mandarin and Spanish are taken into account. Improvement over baseline model (employed in current state-of-the-art automatic accentedness evaluation system) without L1-related features can be observed when adding L1-related features to the input, although the improvement is relatively marginal compared to L1-dependent experiments. It proves that the L1 related information can also help the prediction

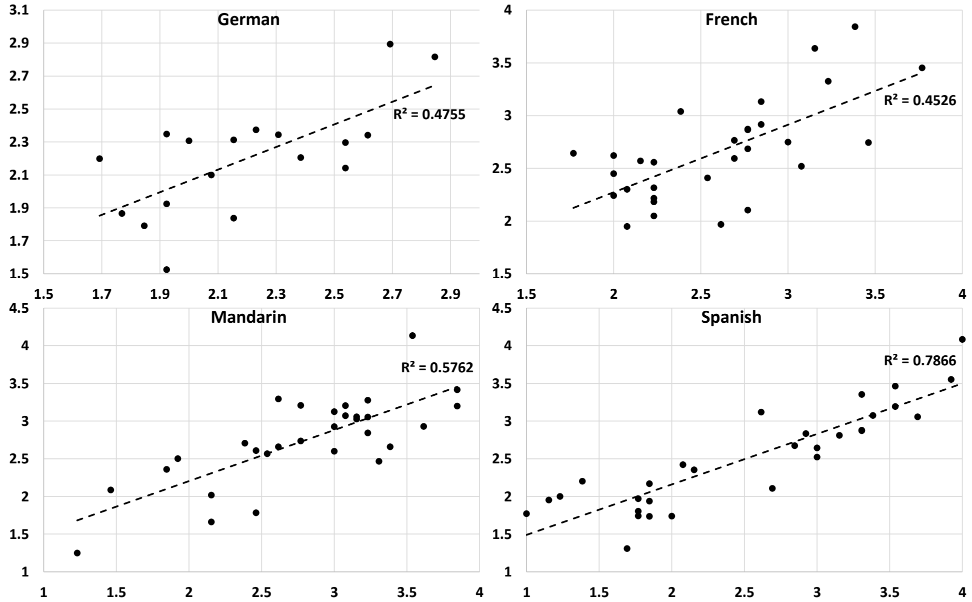


Figure 6.3: Scatter plots of the true labels (X-axis) and predictions (Y-axis) of speakers from different L1s. The fitting line for each L1 is also shown here together with the R-squared values.

in L1-independent case. However, it can be found that the improvement in L1-independent experiment is less than the improvement in L1-dependent experiment. The reason for this will be discussed in next section. Since the improvement is not that large, figure 6.3 only shows the scatter plots achieved with all four feature sets for different L1s. Speakers from different L1s are plot individually with fitting lines and R-squared values. It can be found that Spanish speakers are fitted the best with the highest R-squared value. German and French speakers are not fitted well compared to Spanish and Mandarin speakers. The overall correlation coefficient is 0.811 as shown in figure 6.2, which indicates a strong relationship between model predictions and groundtruth accentedness scores.

6.4 Discussion

This chapter derives a computational model for automatic accentedness evaluation based on findings in previous chapters. The core idea is a new feature extraction scheme that not only quantifies the deviation from native L2 phonological patterns but also how much the accented speech is affected by L1 phonological patterns. Experiments on both L1-dependent and L1-independent tasks show that there is consistent improvement when combining L1 information in the input feature sets.

As a computational model, some blocks of the proposed system can be flexible depending the specific task and resources available. For example, more powerful acoustic models can be used to derive better pronunciation features. More accurate forced-alignment can also be achieved with better forced-alignment tools, thus improving the prosodic features. Depending on the size of dataset, regression models with different complexity can be applied to achieve better performance. Although the evaluation in the current study is done on speaker-level, the proposed framework can be easily extended to sentence-level evaluation.

As mentioned before, the performance improvement of the proposed system on L1-independent tasks is smaller than L1-dependent tasks. This is due to the variability introduced by different L1 acoustic models and different forced-alignment tools for different L1s. This may result in the scales of L1 related features variate for different L1s. This problem can be relieved by using equally powerful L1 acoustic models and forced-alignments, although it could cost much more effort. Another possible solution is to normalize the L1-dependent features with the distance between L1 and L2 pronunciation patterns or the distance between L1 and L2 prosodic patterns. It is not easy to directly calculate the distance between pronunciation patterns of two language given their acoustic models. This study tried to normalize the L1 prosodic measurements with the distance between prosodic patterns of L1 and L2, which almost does not change the final results. This is possibly because there is too much variation of the forced-alignment quality of different L1s. However, this study still believe this is a direction worth more investigation.

Chapter 7

GENERAL DISCUSSION

7.1 Introduction

This chapter provides a general discussion about the experiments and findings in this study. Specially, both theoretic and practical implications of the current study will be introduced. The first section will focus on how the current study contributes to L2 speech learning theories, and the second section will focus on how the current study contributes to automatic accentedness evaluation.

7.2 Implication for L2 speech learning theories

As reviewed in chapter 2, a bunch of studies have investigate the effect of L1 in the acquisition of L2 phonological properties. This effect presents in both segmental (as shown by Strange *et al.* (1992); Flege (1987); Chang *et al.* (2008); Munro (1993); Derakhshan and Karimi (2015)) and suprasegmental acquisition (as shown by Mennen (2004); Stockmal *et al.* (2005); White and Mattys (2007); Lin and Wang (2008); Li and Post (2014); Ordin and Polyanskaya (2015)). However, for segmental properties acquisition, almost all previously mentioned studies have several limitations to comprehensively reveal the detail about the L1's effect on L2 acquisition:

1. Almost all studies only focus on a specific phonological phenomenon, and analyze how L1 affect the production in L2.

2. Usually the numbers of analyzed speakers and L1s are quite limited.

3. Those studies can only show the L1's effect exist, but there is no way to quantize the influence of L1.

4. Few studies investigate the relationship between L1's effect and degree of foreign accent.

For suprasegmental properties acquisition, thanks to the study by Ramus *et al.* (1999); Grabe and Low (2002), several publications use durational rhythmic measurements to show the change of those measurements at different stage of L2 learning. However, these studies still have the similar limitations that the numbers of L1s and speakers per L1 are small. Moreover, no quantified speakers' accentedness scores are available in those studies, only some qualitative ranges (for example, from beginners to advanced learners).

Different from the methodology presented in previous work, the current study proposes to use a computational framework to quantify the L2 speech learning outcomes of tens of speakers from multiple L1s. Both segmental and suprasegmental phonology acquisition are investigated. The analyses done in this study further validate that the influence of L1 exists in both segmental and suprasegmental phonology acquisition during L2 learning. More importantly, with the L1s' information, multiple regression analysis reveals that the accentedness can be better explained. Besides this, the current study further shows that the difference originates from speakers' L1s will also be presented in their accented speech through the analysis of relative importance of segmental and suprasegmental features to accentedness scores. In the following, Mandarin speakers will be taken as examples to illustrate how the methodology proposed by this study can be further utilized to investigate the L2 speech learning process.

Figure 7.1 shows the average pronunciation scores of vowels in accented speech by Mandarin speakers using both L2 (X-axis) and L1 (Y-axis) acoustic models. Each

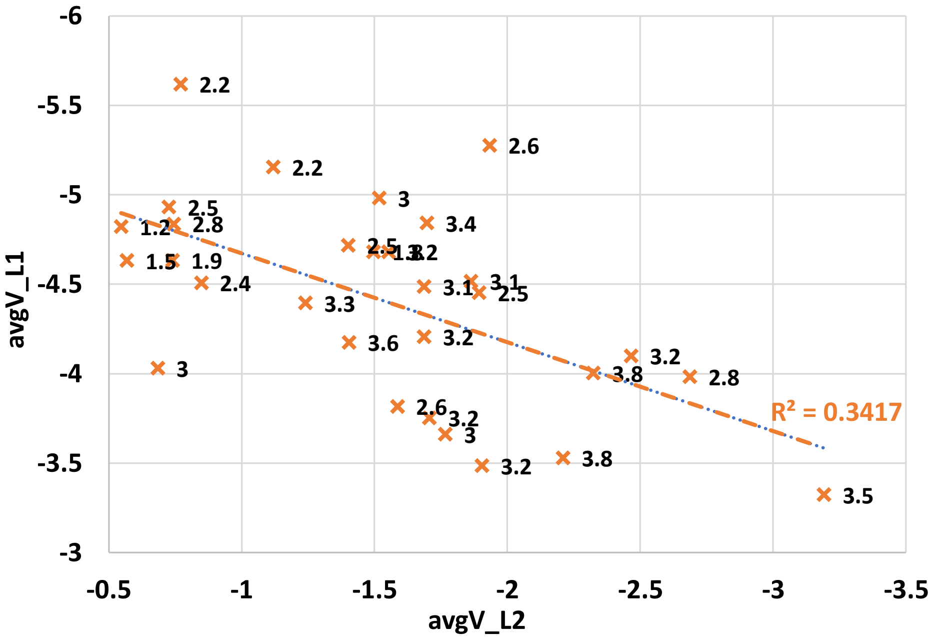


Figure 7.1: The average pronunciation scores of vowels in accented speech by Mandarin speakers using both L2 (X-axis) and L1 (Y-axis) acoustic models. Larger pronunciation score means closer vowel pronunciation to the pronunciation pattern defined by corresponding acoustic model.

speaker (a cross in the figure) has an average vocalic pronunciation score calculated from L2 acoustic model (avgV_L2), and the other one (avgV_L1) is calculated from L1 acoustic model. Larger pronunciation score means closer vowel pronunciation to the pronunciation pattern defined by corresponding acoustic models. The accentedness score of each speaker is also shown along with the crossing on the scatter plot. As shown in figure 4.2, the avgV_L2 has a negative correlation with accentedness score while avgV_L1 has a positive correlation. In order to better show the L1's effect on L2 pronunciation for speakers at different positions on the accentedness scale, this figure plots how similar each speaker's L2 pronunciation is with native L2 and native L1, and together with the accentedness scores. There are several interesting findings in the figure. First, the orange dash line demonstrates the general trend that if one speaker's L2 pronunciation is closer to native L2 speaker, his avgV_L1 score will be lower (means further from L1 pronunciation, thus less affected by L1 phonetic patterns). Second, it can be found that very accented speakers are at the lower-right corner while mildly accented speakers are at the upper-left corner. However, pronunciation can not explain all the variations of accentedness, as indicated by some outliers. For example, two speakers (one with 2.8 accentedness score and the other 3.0) has good pronunciation but are still considered to have strong accented. Third, since the pronunciation score is calculated as the similarity between accented speech and native speech, the positions of native L1 and L2 can not be put on this scatter plot. The L2 can be considered to have 0 avgVL2 value but the avgV_L1 value can not be decided using the current computational model (similar for L1). However, given enough number of speakers, the distance between L1 and L2 can be approximated by the avgV_L1 values of speakers with mildest accent. Fourth, another observation is that there are some obvious outliers which are not right on the transferring path from L1 to L2. For example, both the avgV_L2 (around -1.9) and avgV_L1 (around -5.3) values of the speaker with 2.6 accentedness score 2.6 are relatively low. Those outliers can be attributed to the universal effects mentioned in previous studies (as reviewed by White (1989)), which claim a learner's L2 system have traits that are neither related to L1 nor L2. Major (1987b) also found that the amount of L1's influence decreased as learners become more proficient in L2, and this behavior may vary for different learners.

Figure 7.2 demonstrates the scatter plot between two speech rhythmic measurements: the percentage of vocalic and consonantal durations extracted from Mandarin speakers. These two measurements are chosen because they show more L1's effect for strong accented speakers. X-axis is the percentage of vocalic duration (perV) and Y-axis is the percentage of consonantal duration (perC). Since these measurements are absolute values, both the values of L1, L2 and accented speech can be calculated independently. In the figure, the blue diamond is the position of native English; the

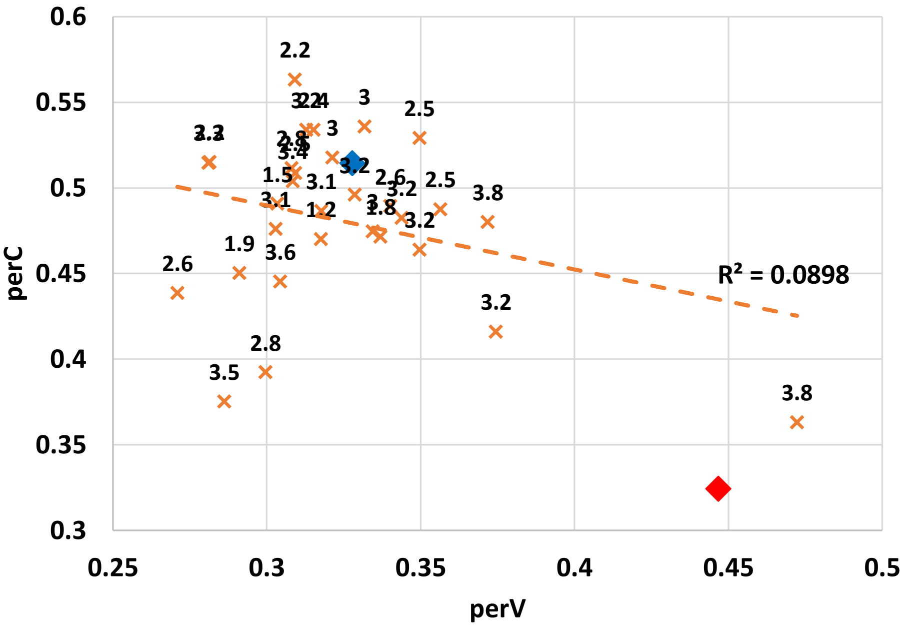


Figure 7.2: Scatter plot of two rhythmic measurements of accented speech by Mandarin speakers: percentage of vocalic (X-axis) and consonantal (Y-axis) durations. The measurements of native English (Blue diamond) and Mandarin (Red diamond) are also shown.

red diamond is the position of native Mandarin; the orange crossings are accented speakers. The trend line (orange dash line) and R-square value are on the accented speakers only. Compared to English, Mandarin has high perV value but lower perC value. It can be found that measurements of most of accented speakers are around the native English, and only part of them are on the path from native L1 to native L2. This observation is in line with previous studies on L2 speech rhythm acquisition (Stockmal *et al.*, 2005; Lin and Wang, 2008; Li and Post, 2014) where the authors show evidences that speech rhythmic measurements are not on the path from L1 to L2, indicating existence of effects that are independent from L1. However, the speaker with the highest accentedness score (3.8) clearly uses the L1 patterns to pronounce English. The results suggests that the prosodic patterns of accented speakers can be affected by L1, but may only influence a few prosodic dimensions or even be independent from L1; some speakers may not be affected by L1 prosodic patterns when producing L2 speech; speakers with mild accent also show no sign of being affected by L1 prosodic patterns.

To summarize, besides the conclusions drew by this study, it is expected that the methodology used in this study could facilitate further research directions on the interference of L1 in L2 speech learning process in a larger scale than previous studies. It can potentially reveal different factors that contribute to the perceived accentedness other than L1's effect. It can also benefit the L2 education field by individually giving a quantitative approximation of the process of L2 speech learning, and providing detailed feedback on which part of the English phonology the learners should focus on in following studies.

7.3 Implication for practical computational models for speech applications

Besides theoretical implications, this study can also contribute to the study on automatic accentedness evaluation. Automatic accentedness (or nativeness) evaluation plays an important role in computer-assisted pronunciation training (CAPT) and computer-aided language learning (CALL). State-of-the-art automatic system includes both the segmental and suprasegmental speech features to model the perception of foreign accent. However, they ignore the effect of L1 in L2 speech learning, and thus can be improved with the computational model proposed in this study. As already shown in chapter 6, adding the contrastive information between accented speech and L1 can improve the performance on accentedness prediction.

Another field that could benefit from the current study is speech intelligibility evaluation of pathological speech. This field is emerging as another important application area of speech technologies with the developing of telemedicine and increasing population impacted by speech disoreders. Although there is great interest in developing computational models for this application, current studies usually develops a feature extraction scheme or directly use existed feature extraction scheme such as Opensmile (Eyben *et al.*, 2010), and then build a machine learning model on the features as presented in my previous studies (Tu *et al.*, 2016a, 2017a,b). The limitation is that existing feature extraction schemes for pathological speech only focus on low-level acoustic features directly calculated on time or frequency domain of original speech signal. This may be suboptimal when the machine learning model is not powerful enough or the amount of data is limited. As shown by Tu *et al.* (2016b), the performance of ASR have very strong correlation with the overall intelligibility of pathological speech. Thus, the computational model proposed in this study (without L1, and replace accented speech with pathological speech and accentedness with intelligibility or severity) can also be used for automatic evaluation of pathological speech.

However, a concern is that whether it is easy to obtain those L1 related features considering the need for a L1 acoustic model. In this study, L1 acoustic models trained on tens of or even over a hundred hours of speech recordings are employed. These L1 datasets may not be available in practice, especially for L1s with small amount of resources (Gales *et al.*, 2017). Indeed, the pronunciation scores based features require acoustic models, but the acoustic models can be trained on small amount of data with phonemes as HMM modeling unit, thus reduce the model space and required training data. There is no need for large vocabulary ASR which usually requires much more speech data. Also, with a simpler acoustic model, the performance of forced-alignment will not be affected very much given known transcription of the accented speech.

Chapter 8

CONCLUSION

8.1 Main findings

This dissertation has investigated the L1's effect on L2 speech learning outcomes using a computational model to analyze accented speech. Motivated by previous findings that L1 can influence phonological system of accented speech in L2, the computational model proposed in this study further validate the statement in a quantitative way by showing how similar the phonological system is with the speaker's L1 phonology. This is achieved by analyzing accented speech in both segmental and suprasegmental feature space macroscopically instead of only looking at one specific phonological phenomenon. Specially, for segmental features, a system for calculating pronunciation scores of phonemes in accented speech from both L1 and L2 acoustic models is proposed to study the pronunciation patterns of accented speech in terms of vowels, consonants and syllables, and compare them to the patterns of native L1 and L2. The pronunciation scores calculated with L1 acoustic model quantify how close the pronunciation of L2 phonemes is to the native pronunciation of the speaker's L1, while the pronunciation scores calculated with l2 acoustic model quantify how close the pronunciation of L2 phonemes is to the native pronunciation of L2. For suprasegmental feature space, speech rhythmic measurements based on durations of vowels, consonants and syllables are calculated by automatic forced-alignment on accented speech. The patterns of native L1 and L2 are also obtained by applying same algorithms on native L1 and L2 speech. Contrastive analysis is done between rhythmic measurements of L1 and accented speech, and L2 and accented speech to quantify the similarity between L1 and accented speech, and L2 and accented speech. Correlation analysis and multiple regression analyses have been conducted on an accented speech dataset consisting of four L1s and 30 speakers from each L1. The findings are summarized as following:

1. The overall pronunciation patterns and prosodic patterns of accented speech are affected by L2 learners' L1. On some specific phonological dimensions, the influences of L1 may be significant while on other dimensions the influence may not be significant. The L1's interference has a negative correlation with accentedness, indicating the negative influence of L1 on L2 speech learning. The results also indicate that there may exist some universal effects, which are independent from L1, influencing the formation of phonological system of accented speech. For example, for learners speaking a syllable-timed language, the general trend, which is independent from learners' L1s, is going towards more stress-timed learning outcome. The inaccuracy may comes from other factors, such as the difficulty to master specific prosodic properties. The computational model employed in this study can quantize the influence of L1 on specific phonological properties.

2. Multiple regression analysis on either segmental or suprasegmental feature space shows that adding contrastive information between L1 and accented speech can improve the perception of accentedness. This proves that L1-related information can help explain the variation of accentedness. Selected L1-related features can provide extra information to the perception of accentedness. When applying the proposed computational model to automatic accentedness evaluation system, adding contrastive L1 information can improve the performance of the system.

3. The relative contribution of segmental and suprasegmental features to the per-

ception of foreign accent depends on how different L1 is from L2 on corresponding feature spaces. The methodology used in this study provides a quantitative way to show the relative importance of segmental and suprasegmental inaccuracy to the perception of foreign accent.

8.2 Future work

There is extra work can be done to improve the accuracy of the computational model used in this study:

1. As mentioned in the dissertation, the accuracy of forced-alignment may affect the accuracy of prosodic measurements. An acoustic model with better performance on accented speech can achieve this.

2. There should be similar scales for L1-related features extracting from different L1s. More consistent L1 acoustic models should be used to extract pronunciation scores. A method to normalize the L1-related features should also be investigated to further improve the performance on automatic accentedness evaluation when there are speakers from multiple L1s.

3. When preparing accented speech dataset, a better control of the distribution of accentedness, which means similar number of speakers at different proficiency level, can further improve the persuasiveness of the results.

There are several interesting directions based on the current study that deserve further investigation:

1. The current study only looks at the overall pronunciation scores of vowels, consonants and syllables. Further investigation on specific phonemes can be done to reveal the L1's effect on specific phonemes, especially for those phonemes that are close to or different from specific L2 phonemes.

2. This study uses speech rhythmic measurements as proxy of speech prosody. Actually, speech prosody includes other factors such as intonation, stress, tempo and pause. Analysis on those prosodic features can result in more comprehensive understanding of L1's effect on L2 speech prosody acquisition.

3. More studies on the amount of L1's effect and universal effects should be done to figure out when and where L1's effect plays a role and when and where universal effects play a role.

4. Applying the methodology to pathological speech is also very intriguing. It can facilitate the study of pathological speech and disease's impact on both segmental and suprasegmental speech features.

REFERENCES

Al-Tamimi, J.-E. and E. Ferragne, "Does vowel space size depend on language vowel inventories? evidence from two arabic dialects and french", in "Ninth European Conference on Speech Communication and Technology", (2005).

Altmann, H., *The perception and production of second language stress: A cross-linguistic experimental study* (University of Delaware Newark, DE, USA, 2006).

Angkititrakul, P. and J. H. Hansen, "Advances in phone-based modeling for automatic accent classification", IEEE Transactions on Audio, Speech, and Language Processing **14**, 2, 634–646 (2006).

Arslan, L. M. and J. H. Hansen, "A study of temporal features and frequency characteristics in american english foreign accent", The Journal of the Acoustical Society of America **102**, 1, 28–40 (1997).

Bernstein, J., M. Cohen, H. Murveit, D. Rtischev and M. Weintraub, "Automatic evaluation and training in english pronunciation.", in "ICSLP", vol. 90, pp. 1185–1188 (1990).

Best, C. T., "Chapter6 aˆ a direct realist view of cross-language speech perception", Speech perception and linguistic experience: Issues in cross-language research pp. 171–204 (1995).

Best, C. T. and M. D. Tyler, "Nonnative and second-language speech perception: Commonalities and complementarities", Language experience in second language speech learning: In honor of James Emil Flege **1334**, 1–47 (2007).

Bickel, B., "What is typology?a short note", unpublished paper, University of Leipzig (2001).

Black, M. P., D. Bone, Z. I. Skordilis, R. Gupta, W. Xia, P. Papadopoulos, S. N. Chakravarthula, B. Xiao, M. Van Segbroeck, J. Kim *et al.*, "Automated evaluation of non-native english pronunciation quality: combining knowledge-and data-driven features at multiple time scales.", in "INTERSPEECH", pp. 493–497 (2015).

Bradlow, A. R., "A comparative acoustic study of english and spanish vowels", The Journal of the Acoustical Society of America **97**, 3, 1916–1924 (1995).

Bu, H., J. Du, X. Na, B. Wu and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline", arXiv preprint arXiv:1709.05522 (2017).

Butcher, A., "The influence of the native language on the perception of vowel quality", Arbeitsberichte Kiel , 2, 1–137 (1978).

Chang, C., E. Haynes, Y. Yao and R. Rhodes, "The phonetic space of phonological categories in heritage speakers of mandarin", in "Proceedings from the Annual Meeting of the Chicago Linguistic Society", vol. 44, pp. 31–45 (Chicago Linguistic Society, 2008).

Chang, C. B., *First language phonetic drift during second language acquisition* (University of California, Berkeley, 2010).

Chen, L.-Y. and J.-S. R. Jang, “Automatic pronunciation scoring with score combination by learning to rank and class-normalized dp-based quantization”, IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) **23**, 11, 1737–1749 (2015).

Choueiter, G., G. Zweig and P. Nguyen, “An empirical study of automatic accent classification”, in “Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on”, pp. 4265–4268 (IEEE, 2008).

Dellwo, V., “Rhythm and speech rate: A variation coefficient for δc”, Language and language-processing pp. 231–241 (2006).

Derakhshan, A. and E. Karimi, “The interference of first language and second language acquisition”, Theory and Practice in Language Studies **5**, 10, 2112 (2015).

Deterding, D., “The pronunciation of english by speakers from china”, English World-Wide **27**, 2, 175–198 (2006).

Dryer, M. S. and M. Haspelmath, eds., *WALS Online* (Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013), URL `http://wals.info/`.

Eckman, F. R., “Markedness and the contrastive analysis hypothesis”, Language learning **27**, 2, 315–330 (1977).

Elisseeff, A., M. Pontil *et al.*, “Leave-one-out error and stability of learning algorithms with applications”, NATO science series sub series iii computer and systems sciences **190**, 111–130 (2003).

Eyben, F., M. Wöllmer and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor”, in “Proceedings of the 18th ACM international conference on Multimedia”, pp. 1459–1462 (ACM, 2010).

Fant, G., “Speech sounds and features.”, (1973).

Flege, J. and O. Bohn, “Perception and production of a new vowel category by adult second language learners”, Second-language speech: Structure and process **13**, 53 (1997).

Flege, J. E., “The production of new and similar phones in a foreign language: Evidence for the effect of equivalence classification”, Journal of phonetics **15**, 1, 47–65 (1987).

Flege, J. E., “The production and perception of foreign language speech sounds”, Human communication and it’s disorders-a review pp. 224–401 (1988).

Flege, J. E., “Second language speech learning: Theory, findings, and problems”, Speech perception and linguistic experience: Issues in cross-language research pp. 233–277 (1995).

Flege, J. E., “English vowel production by dutch talkers: more evidence for the similar vs.new distinction”, Second-language speech pp. 11–52 (1997).

Flege, J. E., I. R. MacKay and D. Meador, “Native italian speakers perception and production of english vowels”, The Journal of the Acoustical Society of America **106**, 5, 2973–2987 (1999).

Flege, J. E., M. J. Munro and I. R. MacKay, “Factors affecting strength of perceived foreign accent in a second language”, The Journal of the Acoustical Society of America **97**, 5, 3125–3134 (1995).

Franco, H., L. Neumeyer, Y. Kim and O. Ronen, “Automatic pronunciation scoring for language instruction”, in “Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on”, vol. 2, pp. 1471–1474 (IEEE, 1997).

Gales, M. J., K. M. Knill and A. Ragni, “Low-resource speech recognition and keyword-spotting”, in “International Conference on Speech and Computer”, pp. 3–19 (Springer, 2017).

Gil, D., “A prosodic typology of language”, Folia Linguistica **20**, 1-2, 165–232 (1986).

Grabe, E. and E. L. Low, “Durational variability in speech and the rhythm class hypothesis”, Papers in laboratory phonology **7**, 515-546 (2002).

Grósz, T., R. Busa-Fekete, G. Gosztolya and L. Tóth, “Assessing the degree of nativeness and parkinson’s condition using gaussian processes and deep rectifier neural networks”, in “INTERSPEECH”, (2015).

Gut, U., *Non-native speech: A corpus-based analysis of phonological and phonetic properties of L2 English and German*, vol. 9 (Peter Lang, 2009).

Hardman, J., “Accentedness and intelligibility of mandarin-accented english for chinese, koreans and americans”, in “Proceedings of the International Symposium on the Acquisition of Second Language Speech”, (2014).

Jongman, A., M. Fourakis and J. A. Sereno, “The acoustic vowel space of modern greek and german”, Language and speech **32**, 3, 221–248 (1989).

Kang, O., “Relative salience of suprasegmental features on judgments of l2 comprehensibility and accentedness”, System **38**, 2, 301–315 (2010).

Kawase, S., J. Kim and C. Davis, “The influence of second language experience on japanese-accented english rhythm”, Proceeding of Speech Prosody 2016 (2016).

Kim, Y., H. Franco and L. Neumeyer, “Automatic pronunciation scoring of specific phone segments for language instruction.”, in “Eurospeech”, (1997).

Kunath, S. A. and S. H. Weinberger, “The wisdom of the crowd’s ear: speech accent rating and annotation with amazon mechanical turk”, in “Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon’s Mechanical Turk”, pp. 168–171 (Association for Computational Linguistics, 2010).

Lai, C., K. Evanini and K. Zechner, “Applying rhythm metrics to non-native spontaneous speech.”, in “SLaTE”, pp. 159–163 (2013).

Li, A. and B. Post, “L2 acquisition of prosodic properties of speech rhythm: Evidence from l1 mandarin and german learners of english”, Studies in Second Language Acquisition **36**, 2, 223–255 (2014).

Lin, H. and Q. Wang, “Interlanguage rhythm in the english production of mandarin speakers”, in “8th Phonetic Conference of China and the InternationalSymposium of Phonetic Frontiers”, (2008).

Lisker, L. and A. S. Abramson, “A cross-language study of voicing in initial stops: Acoustical measurements”, Word **20**, 3, 384–422 (1964).

Littel, P., D. R. Mortensen and L. Levin, “Uriel typological database”, Pittsburgh: CMU (2016).

Long, M. H., “Maturational constraints on language development”, Studies in second language acquisition **12**, 03, 251–285 (1990).

Maddieson, I., “Ucla phonological segment inventory database, version 1.1”, Los Angeles: Dept. of Linguistics, UCLA (1992).

Major, R. C., “English voiceless stop production by speakers of brazilian portuguese”, Journal of Phonetics **15**, 2, 197–202 (1987a).

Major, R. C., “A model for interlanguage phonology”, Interlanguage phonology: The acquisition of a second language sound system pp. 101–124 (1987b).

McAuliffe, M., M. Socolof, S. Mihuc, M. Wagner and M. Sonderegger, “Montreal forced aligner: trainable text-speech alignment using kaldi”, in “Proceedings of interspeech”, (2017).

McCullough, E. A., *Acoustic correlates of perceived foreign accent in non-native English*, Ph.D. thesis, The Ohio State University (2013a).

McCullough, E. A., “Perceived foreign accent in three varieties of non-native english”, Ohio State U. Working Papers in Linguistics **60**, 51–66 (2013b).

Mennen, I., “Bi-directional interference in the intonation of dutch speakers of greek”, Journal of phonetics **32**, 4, 543–563 (2004).

Mok, P. P. and V. Dellwo, “Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, beijing mandarin and english”, in “Proceedings of Speech Prosody”, vol. 4 (Citeseer, 2008).

Moran, S., D. McCloy and R. Wright, eds., *PHOIBLE Online* (Max Planck Institute for Evolutionary Anthropology, Leipzig, 2014), URL `http://phoible.org/`.

Moravcsik, E. A., *Introducing language typology* (Cambridge University Press, 2012a).

Moravcsik, E. A., *What is language typology?*, p. 1C24, Cambridge Introductions to Language and Linguistics (Cambridge University Press, 2012b).

Moustroufas, N. and V. Digalakis, "Automatic pronunciation evaluation of foreign speakers using unknown text", Computer Speech & Language **21**, 1, 219–230 (2007).

Munro, M. J., "Productions of english vowels by native speakers of arabic: Acoustic measurements and accentedness ratings", Language and Speech **36**, 1, 39–66 (1993).

Munro, M. J. and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners", Language learning **45**, 1, 73–97 (1995).

Munro, M. J. and T. M. Derwing, "Modeling perceptions of the accentedness and comprehensibility of l2 speech the role of speaking rate", Studies in second language acquisition **23**, 04, 451–468 (2001).

Munro, M. J., T. M. Derwing and C. S. Burgess, "Detection of nonnative speaker status from content-masked speech", Speech communication **52**, 7, 626–637 (2010).

Munro, M. J., J. E. Flege and I. R. MacKay, "The effects of age of second language learning on the production of english vowels", Applied psycholinguistics **17**, 3, 313–334 (1996).

Neri, A., O. Mich, M. Gerosa and D. Giuliani, "The effectiveness of computer assisted pronunciation training for foreign language learning by children", Computer Assisted Language Learning **21**, 5, 393–408 (2008).

Neumeyer, L., H. Franco, M. Weintraub and P. Price, "Automatic text-independent pronunciation scoring of foreign language student speech", in "Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on", vol. 3, pp. 1457–1460 (IEEE, 1996).

Ordin, M. and L. Polyanskaya, "Acquisition of speech rhythm in a second language by learners with rhythmically different native languages", The Journal of the Acoustical Society of America **138**, 2, 533–544 (2015).

Panayotov, V., G. Chen, D. Povey and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books", in "Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on", pp. 5206–5210 (IEEE, 2015).

Parmenter, C. E. and A. V. Blanc, "An experimental study of accent in french and english", Publications of the Modern Language Association of America pp. 598–607 (1933).

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, "Scikit-learn: Machine learning in Python", Journal of Machine Learning Research **12**, 2825–2830 (2011).

Polyanskaya, L., M. Ordin and M. G. Busa, “Relative salience of speech rhythm and speech rate on perceived foreign accent in a second language”, Language and Speech p. 0023830916648720 (2016).

Povey, D., A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, “The kaldi speech recognition toolkit”, in “IEEE 2011 workshop on automatic speech recognition and understanding”, No. EPFL-CONF-192584 (IEEE Signal Processing Society, 2011).

Qian, Y., K. Evanini, X. Wang, C. M. Lee and M. Mulholland, “Bidirectional lstm-rnn for improving automated assessment of non-native children?s speech”, Proc. Interspeech 2017 pp. 1417–1421 (2017).

Raab, M., R. Gruhn and E. Noeth, “Non-native speech databases”, in “Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on”, pp. 413–418 (IEEE, 2007).

Rabiner, L. R., “A tutorial on hidden markov models and selected applications in speech recognition”, Proceedings of the IEEE **77**, 2, 257–286 (1989).

Ramus, F., M. Nespor and J. Mehler, “Correlates of linguistic rhythm in the speech signal”, Cognition **73**, 3, 265–292 (1999).

Rasier, L. and P. Hiligsmann, “Prosodic transfer from l1 to l2. theoretical and methodological issues”, Nouveaux cahiers de linguistique française **28**, 2007, 41–66 (2007).

Rasipuram, R., M. Cernak, A. Nanchen *et al.*, “Automatic accentedness evaluation of non-native speech using phonetic and sub-phonetic posterior probabilities”, in “Proceedings of Interspeech”, No. EPFL-CONF-209089 (2015).

Ribeiro, E., J. Ferreira, J. Olcoz, A. Abad, H. Moniz, F. Batista and I. Trancoso, “Combining multiple approaches to predict the degree of nativeness.”, in “INTERSPEECH”, pp. 488–492 (2015).

Rognoni, L. and M. G. Busà, “Testing the effects of segmental and suprasegmental phonetic cues in foreign accent rating: An experiment using prosody transplantation”, in “Proc. International Symposium on the Acquisition of Second Language Speech”, pp. 547–560 (2013).

Rzeszotarski, J. M., E. Chi, P. Paritosh and P. Dai, “Inserting micro-breaks into crowdsourcing workflows”, in “First AAAI Conference on Human Computation and Crowdsourcing”, (2013).

Saeys, Y., I. Inza and P. Larrañaga, “A review of feature selection techniques in bioinformatics”, bioinformatics **23**, 19, 2507–2517 (2007).

Saito, K., P. Trofimovich and T. Isaacs, “Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels”, Applied Psycholinguistics **37**, 2, 217–240 (2016).

Sangwan, A. and J. H. Hansen, "Automatic analysis of mandarin accented english using phonological features", Speech Communication **54**, 1, 40–54 (2012).

Schuller, B. W., S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang and F. Weninger, "The interspeech 2015 computational paralinguistics challenge: nativeness, parkinson's & eating condition.", (2015).

Shah, A., "Temporal characteristics of spanish-accented english: Acoustic measures and their correlation with accentedness ratings", Unpublished Dissertation, The City University of New York, New York City (2002).

Sheldon, A. and W. Strange, "The acquisition of/r/and/l/by japanese learners of english: Evidence that speech production can precede speech perception", Applied psycholinguistics **3**, 3, 243–261 (1982).

Snow, R., B. O'Connor, D. Jurafsky and A. Y. Ng, "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks", in "Proceedings of the conference on empirical methods in natural language processing", pp. 254–263 (Association for Computational Linguistics, 2008).

Stockmal, V., D. Markus and D. Bond, "Measures of native and non-native rhythm in a quantity language", Language and speech **48**, 1, 55–63 (2005).

Strange, W., E. Tohkura, E. Vatikiotis-Bateson and Y. Sagisaka, "Learning non-native phoneme contrasts: Interactions among subject, stimulus, and task variables", Speech perception, production and linguistic structure pp. 197–219 (1992).

Tao, J., S. Ghaffarzadegan, L. Chen and K. Zechner, "Exploring deep learning architectures for automatically grading non-native spontaneous speech", in "Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on", pp. 6140–6144 (IEEE, 2016).

Tu, M., V. Berisha and J. Liss, "Interpretable objective assessment of dysarthric speech based on deep neural networks", Proc. Interspeech 2017 pp. 1849–1853 (2017a).

Tu, M., V. Berisha and J. Liss, "Objective assessment of pathological speech using distribution regression", in "Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on", pp. 5050–5054 (IEEE, 2017b).

Tu, M., A. Grabek, J. Liss and V. Berisha, "Investigating the role of l1 in automatic pronunciation evaluation of l2 speech", in "Proceedings of Interspeech 2018", (to be published, 2018).

Tu, M., Y. Jiao, V. Berisha and J. M. Liss, "Models for objective evaluation of dysarthric speech from data annotated by multiple listeners", in "Signals, Systems and Computers, 2016 50th Asilomar Conference on", pp. 827–830 (IEEE, 2016a).

Tu, M., A. Wisler, V. Berisha and J. M. Liss, "The relationship between perceptual disturbances in dysarthric speech and automatic speech recognition performance", The Journal of the Acoustical Society of America **140**, 5, EL416–EL422 (2016b).

van Maastricht, L., T. Zee, E. Krahmer and M. Swerts, “L1 perceptions of l2 prosody: The interplay between intonation, rhythm, and speech rate and their contribution to accentedness and comprehensibility”, Proc. Interspeech 2017 pp. 364–368 (2017).

Wagner, A. and A. Braun, “Is voice quality language-dependent? acoustic analyses based on speakers of three different languages”, Language **6**, 4, 2 (2003).

Weiher, E., “Lautwahrnehmung und lautproduktion im englischunterricht für deutsche.(perception et production des sons dans l’enseignement de l’anglais à des allemands)”, Arbeitsberichte Kiel , 3, 1–243 (1975).

Weinberger, S., “Speech accent archive”, George Mason University (2013).

White, L., *Universal grammar and second language acquisition*, vol. 1 (John Benjamins Publishing, 1989).

White, L. and S. L. Mattys, “Calibrating rhythm: First language and second language studies”, Journal of Phonetics **35**, 4, 501–522 (2007).

William, F., A. Sangwan and J. H. Hansen, “Automatic accent assessment using phonetic mismatch and human perception”, IEEE transactions on audio, speech, and language processing **21**, 9, 1818–1829 (2013).

Winters, S. and M. G. O’Brien, “Perceived accentedness and intelligibility: The relative contributions of f0 and duration”, Speech Communication **55**, 3, 486–507 (2013).

Witt, S. M. and S. J. Young, “Phone-level pronunciation scoring and assessment for interactive language learning”, Speech communication **30**, 2-3, 95–108 (2000).

Xi, X., H. Franco, H. Bratt, R. Rossier, V. Rao Gadde, E. Shriberg, V. Abrash and K. Precoda, “Eduspeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications”, Language Testing **27**, 3, 401–418 (2010).

Yuan, J., Y. Jiang and Z. Song, “Perception of foreign accent in spontaneous l2 english speech”, in “Speech Prosody 2010-Fifth International Conference”, (2010).

APPENDIX A

INSTRUCTIONS AND TASK INFORMATION FOR ACCENTEDNESS SCORE COLLECTION ON AMT

1. Please do not use Back/Refresh buttons during this task.

2. In case you want to end, close the tab. To resume, log back in.

3. This task is about the degree of accentedness of the speaker speaking English. Please focus on how different the speaker sounds from a native speaker of American English in the pronunciation of sounds and words, stress and intonation position, the way to combine different sounds and words into a sentence.

4. You will be asked to give your general impression of the speakers degree of accentedness on a 1-4 scale (1 for negligible/no accent, 2 for mild accent, 3 for strong accent and 4 for very strong accent) and whether you are certain about your answer (certain or uncertain). There will be four examples before the listening task for you to better understand the degree of accentedness.

5. There are 150 audio files in this task, each of which is 10 seconds. This task will take about 40 minutes.

6. You are allowed to listen to each sentence twice.

7. Please find a quiet place to perform this task.

8. We recommend using Chrome for this task.

9. Go to the URL, create an account (no pw required), and complete the task. Remember your created username and write it down in the textbox below. This is IMPORTANT because we'll use it to link your work with your MTurk account so that we can pay you if approved. Note that you have to COMPLETE the whole task to receive the reward, and we'll only take ONE completed task from each participant. So please enter only one username in the textbox below.

10. Please do the task in the survey link URL.

APPENDIX B

CONSENT FORM FOR ACCENTEDNESS SCORE COLLECTION ON AMT

Introduction

The purposes of this form are to provide you (as a prospective research study participant) information that may affect your decision as to whether or not to participate in this research and to record the consent of those who agree to be involved in the study.

Researchers

Dr. Julie Liss, a Professor in the Department of Speech & Hearing Sciences (College of Health Solutions) at ASU, and Dr. Visar Berisha, an Assistant Professor in the Department of Speech & Hearing Sciences and the School of Electrical, Computer, and Energy Engineering at ASU, have invited your participation in a research study.

Study purpose

We are collecting perceived degree of accentedness from people aged 18 and older who have normal hearing. We will use these accentedness ratings to study the impact of non-native English speakers native language on the perceived accentedness.

Description of research study

If you decide to participate, then you will join a study involving research of the perception of accented speech. Your participation will be completely online and will last no longer than 1 hour. If you agree to participate, we ask that you be seated in a quiet room in front of a computer. You will listen to a paragraph spoken by different individuals in English and asked to give a general impression of the accentedness of each speaker on a 1-4 scale. Research completed based on these accentedness ratings will provide an understanding of the impact of non-native English speakers native language on perceived accentedness.

Risks

There are no known risks from taking part in this study.

Benefits

Although there may be no direct benefits to you, these transcriptions may improve our understanding of accented speech. This may, in turn, allow for the development of computer-aided second language learning system.

Confidentiality

All information obtained in this study is strictly confidential. The results of this research study may be used in reports, presentations, and publications, but the researchers will not identify you.

Withdraw privilege

Your participation in this project is completely voluntary. There is no penalty for not participating, or for choosing to withdraw from participation at any time. Your decision will in no way affect your relationship with ASU or your grade in any course. Should you choose to withdraw from the study, your digital audio-video files will not be saved and will be discarded electronically.

Costs and payments

The researchers want your decision about participating in the study to be absolutely voluntary. Yet they recognize that your participation may pose some inconvenience. You will receive $1.5 for your participation, paid via Amazon Mechanical Turk.

Voluntary consent

Any questions you have concerning the research study or your participation in the study, before or after your consent, will be answered by Dr. Julie Liss at (480) 965-9136. If you have questions about your rights as a subject/participant in this research, or if you feel you have been placed at risk; you can contact the Chair of the Human Subjects Institutional Review Board, through the ASU Office of Research Integrity and Assurance, at 480-965 6788. This form explains the nature, demands, benefits and any risk of the project. By signing this form you agree knowingly to assume any risks involved. Remember, your participation is voluntary. You may choose not to participate or to withdraw your consent and discontinue participation at any time without penalty or loss of benefit. In signing this consent form, you are not waiving any legal claims, rights, or remedies. A copy of this consent form will be offered to you.

By clicking “Agree”, you consent to participate in the above study and indicated that:

1. you have read the above information
2. you voluntarily agree to participate
3. you are at least 18 years of age