UNIVERSITY OF AMSTERDAM

Cross-Demographic Acoustic Analysis of the Labiodental Approximant [ʋ] Sound Change in Putonghua

MA Thesis

University of Amsterdam
Graduate School of Humanities

Submission June 2022

Jordan Crowley 13618482
[Soon Jordan C. Martin]

Supervisor: Dr. Marijn van't Veer

Acknowledgements

I would like to express my gratitude to my adviser Dr. Marijn van't Veer for his effective supervision. We worked together very well, as his personality and advising style made writing this thesis very enjoyable and time efficient.

I also would like the thank Dr. Titia Benders for her immeasurable help with the statistics components of this thesis.

Lastly, to my colleague Lydia for being a friend to share ideas with and my mother for trusting me to move to Amsterdam alone to pursue this degree.

Abstract

Some native speakers of modern standard spoken Chinese (*Pǔtōnghuà*) have previously been shown to replace the salient labiovelar approximant [w] that is found in many of the world's languages with the labiodental approximant [ʋ]. The demographics that have historically been shown to undergo this sound change most are young, female, and live in Northern regions of China. This thesis explores the presence of this sound change in the demographics not typically expected to produce it, specifically male and Southern speakers of *Pǔtōnghuà*. An acoustic analysis was performed on 1,000 sound recordings from a representative sample of the four demographic combinations (Northern female, Northern male, Southern female, Southern male) via Praat script, then underwent statistical analysis via linear mixed-effects regression modeling in R. The measurements recorded via Praat script were $F2_{min}$, F2 slope, and Harmonics-to-Noise ratio, and results were each presented as a function of one of three binary predictors: region (North vs. South), gender (male vs. female), and word (wei vs. wen). The results from $F2_{min}$ were inconclusive for the particular research question of this thesis, but the results for F2 slope and HNR measurements show quite strongly that male speakers may in fact use the labiodental variant [ʋ] more than females, which opposes much of the established literature on this sound change within the broader scope of Chinese linguistics.

Table of Contents

1. Introduction

Language is not always concrete. The well-established pronunciation systems of language are anything but stagnant, even for widely spoken languages such as English and Chinese. It is possible, for whatever reason, that certain pronunciations will change slightly from what they once were; this is called sound change, sometimes phonetic change or phonologization (Hyman 1976: 171). The direct link to a specific language's phonology is beyond the scope of this thesis, but it may be interesting for some readers. The sound change of interest in this thesis is the change from the labiovelar approximant [w] into a slight variant of that, a labiodental approximant [ʋ]. This change has been shown to exist in Chinese, and this thesis will delve into some of the intricacies of how and to what extent this sound change truly exists within the language.

This thesis will heavily rely on acoustics and the connection between acoustics and articulation. With respect to Chinese, this sound change only represents a small piece of the broad jurisdiction of Chinese phonology. Because of this limited scope, there will be discussion of relevant features of Chinese phonology, but the thesis is not focused on how sounds pattern together within Chinese, especially outside of the reticle of labiovelar or labiodental approximants. While by no means an all-encompassing investigation of the sound change under question, my hope is that this thesis fills a substantial niche of knowledge at the intersection between Chinese philology, phonetics, and sociolinguistics.

Finally, though experts on any one subject can come from anywhere, it is important to know when learned knowledge is not enough, and when to amplify the voices of others, whose lived experiences supersede one's own. I enjoy studying Chinese linguistics, as it is fascinating historically and phonologically, and I consider myself to have useful knowledge of both, which aids in my conveying of information as I see fit. In my writing, I try to accurately portray many nuanced linguistic features of this language, but I gladly acknowledge that I will never have the first-hand experience that native researchers do; to that standard, Chinese linguists must have the final say, and I do my best to rely on their literature as much as is possible.

1.1 Research Questions

I want to know how often this sound change occurs in non-Northern regions of China as well as in non-female speakers of *Pǔtōnghuà*. Because the previous literature has well-established that northern females use [ʋ] more often than any other demographic, this felt like a clean subset of data to use a baseline. However, due to the statistical analyses ran for this thesis, it is more accurate to phrase the question and hypotheses in terms of relativistic comparisons between predictors (variables). As a result, the goal of this thesis is to compare the usage of [ʋ] (via indicative acoustic measurements) across three predictors: namely gender, region, and word (§ 6.2) and determine which of these predictors are the most likely to have an influence on the production of this sound change.

The literature is clear that Northerners (defined in § 5.2) and female speakers are expected to use this sound change the most. However, my hypotheses vary from what is long-established in the literature. It has been over a decade since the labiodental approximant [ʋ] has been studied in a similar manner for *Pǔtōnghuà* as in this thesis (Wiener & Shih 2011), so my working hypotheses are as follows:

1. Going against the established literature (Hu 1991, Hu 1987), I expect that the usage of [ʋ] in male speakers may in fact rival its usage by female speakers.
2. Going against the established literature (Shen 1987, Wiener & Shih 2011), I expect that the usage of [ʋ] will be just as prominent in the South as it is in the North.

3. In line with the literature, I continue to expect that the word /wen/ is realized as [ʋen] more than /wei/ is realized as [ʋei].

This thesis does not attempt to answer the question as to what may cause this shift in sound change, only that this shift has occurred. The rationale for why this thesis takes this stance will be elaborated upon in chapter 5.

These questions will be answered via a phonetic analysis of many sound files, obtained via a public corpus, of natural Chinese speech. Analyzing the concrete values of speech ($F2_{min}$, F2 slope, and HNR) is the most effective way to ensure that the results I suggest are truly the result of what I intend to study. § 8.4 discusses some possible flaws with this type of analysis, but primarily, the acoustic analysis is the best objective method available.

1.2 Road Map

This first several sections lay out all the necessary theory and background required to understand the rest of this thesis. It begins with a description and classification on the languages of China, how to properly categorize the different words used for the term *Chinese* as well as the phonological norms and features present in the languages of China. Then, an overview of general acoustics will be covered such as basic wave equations, tube resonators, and the harmonics-to-noise ratio. Following the pure acoustics, characteristics of approximants broadly as well as specific features of the labiovelar approximant [w] as well as the labiodental approximant [ʋ] will be discussed. This includes things such as coarticulation concerns, formant patterns, and how the effects of articulation map onto specific acoustic patterns. Finally, the theory is wrapped up with sociolinguistic and demographic information relevant to the analysis of this thesis. After the theoretical grounding, the roadmap of this paper is straightforward, with the methods, results, discussion, and conclusion.

1.3 Terminology Establishment

A few notations and conventions that will be used in this thesis will now be explained here. First, all transliterations and romanizations of Chinese words will be accomplished using *Pīnyīn*, an explanation of which will be found in § 2.3. Next, the term *Pǔtōnghuà* will be used in most situations of this these instead of the slightly ambiguous terms *Standard Chinese* or *Mandarin*. The completely unambiguous definition of *Pǔtōnghuà* is explained in § 2.1. Finally, all *Pīnyīn* and word to-be-defined will be italicized for standardization and emphasis, while all IPA notations will be contained within [square brackets]. Phonological representations will be shown in /slashes/.

2. Chinese Language and Phonology

This first theory chapter lays out all the necessary background required to understand the different terms of Chinese, the basis of language differences within China as well as the phonological basis of the current version of *Pǔtōnghuà* that will be studied here. It begins with a description and classification on the languages of China, how to properly categorize the different words used for the term *Chinese* as well as the phonological norms and features present in the languages of China.

2.1 History and Development of *Pǔtōnghuà*

The Chinese language is one of the oldest, largest, continually spoken language groups in the world (Ethnologue 1992). Comprising one of the major subgroups of the Sino-Tibetan language family, Chinese encompasses seven modern dialect sub-groups. In decreasing number of native speakers, these seven groups are: Mandarin, Wu, Yue (Cantonese), Min, Kejia (Hakka), Gan, and Xiang (Zhou 2009). Despite all being members of the same large classifying language family, most of these dialect groups have a certain degree of mutual unintelligibility. For example, it is expected that a native speaker of Mandarin is able to understand very little dialogue from a native speaker of Yue or Min. It is also important to note that this dialect breakdown only includes the languages of the majority Han nationality, which make up 92% of all people living in China (Sun 2006). There are many smaller (comparatively) ethnic groups in China, like the Tibetan, Uyghur, Manchu, and Mongolian who have their own language, cultures, and sometimes scripts, but this is not inherently relevant to this thesis. Despite massive amounts of spoken unintelligibility, one unifying feature of the Han languages of China is their identical script. Because of non-sound-based graphemes, all Han languages of China use the same writing system. Despite having (sometimes completely different) pronunciations and phonotactics, there is no diminished extent of Chinese national identity (Norman 1988).

An important note in the classification and description of Chinese languages is the wording used to describe each one. Precise definitions of Mandarin, *Zhōngwén*, *Hànyǔ*, and *Pǔtōnghuà* are necessary. As already described, the term *Chinese* is a large umbrella term, with one of the spoken dialects being *Mandarin*. To be more accurate, *Mandarin* is considered the spoken dialect of *Northern* Chinese, broadly. The exact ramifications of this description will be explained later at the divide between Northern and Southern attempts at language unification. Mandarin is by far the most spoken variety of Chinese, and despite having some small regional variation, the over 700 million native speakers are almost completely intelligible with one another. The term *Zhōngwén* translates literally to *central language*, and is the overall, most accepted and official way to describe *Chinese*. This word is used in most university language departments, government proceedings, and law. If a native Chinese speaker were asked to translate the word *Chinese*, it almost certainly would be *Zhōngwén*. Third, the word *Hànyǔ* describes the languages spoken by the largest ethnic group of China: the Han, and as described previously, the seven subgroups within the umbrella term *Chinese* could be considered *Hànyǔ*, and because of this distinction between Han and non-Han, foreign students learning Chinese are now said to be studying *Hànyǔ* (Sun 2006). Finally, *Pǔtōnghuà* literally means *common language* and refers to the standard version of spoken Chinese that is usable by the entire population, which was intentionally developed and constructed following the fall of the Qing dynasty in 1912. If someone wanted to speak to any other person in China, this would be the dialect to use. Chen (1999: 24) cited the original 1955 *National Conference on Script Reform* definition of *Pǔtōnghuà* as "the standard form of Modern Chinese with the Beijing phonological system as its norm of pronunciation, and Northern dialects as its base dialects."

Why exactly the Northern dialects and Beijing phonological system came to be used as the standard will be explained next. However, with respect to this thesis, all versions of Chinese studied and analyzed will technically be different accents of *Pǔtōnghuà*, as that is the version used by the entire country, and the standard that is taught in schools. In that regard, simply saying *Mandarin* may technically pose a degree of inaccuracy.

2.1.1. Phonological Basis of *Pǔtōnghuà*

To understand why the Beijing dialect became the phonological basis of the standard of Chinese, some history needs to be known; this is not a history thesis, so only important points will be addressed. In 1912, the prolific Qing Dynasty fell to massive uprisings of a new Nationalist Party. One goal of this new party was to create national identity through a unified language and new pronunciation. An attempt was made by scholars of the time to create a pronunciation system representative of the entire country, pulling features from all corners of the nation. However, this caused the new pronunciation system to use some sounds that could not be produced by speakers of Northern dialects and some sounds that could not be produced by speakers of Southern dialects. In their attempt to please everybody, they pleased nobody. At one point, some prominent Chinese scholars considered completely abandoning the Chinese language for Esperanto (Ramsey 1987) in the name of unification.

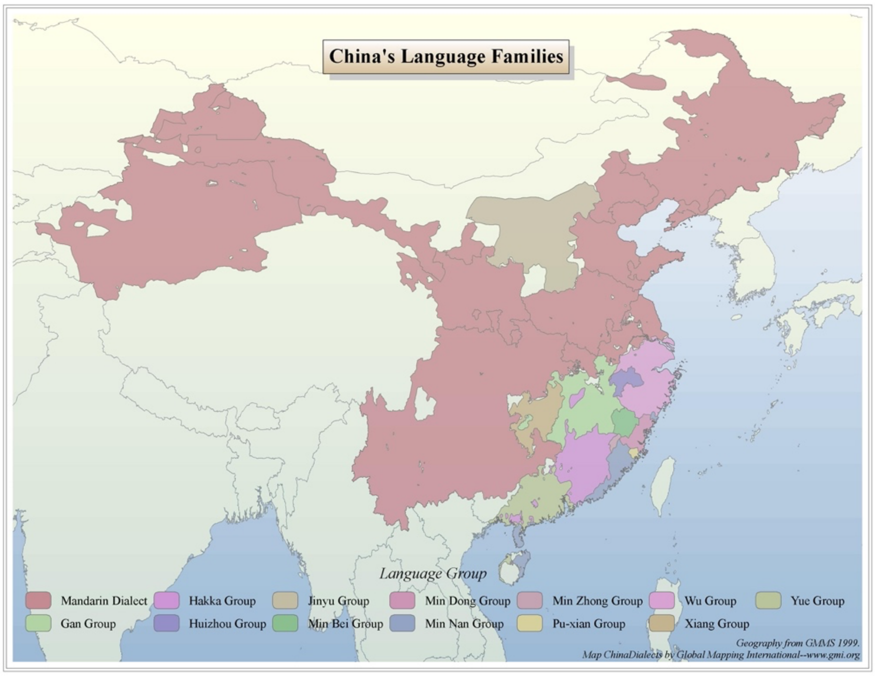


Fig 2.1: Map of language regions and families within mainland China

For nearly 20 years, there was a fragmented standard national language that few people could speak or understand. In 1932, this attempt at an amalgamized language was abandoned and instead replaced with a new one based entirely from the Beijing phonological system. Throughout China's history, the geography of the land made living conditions better and communication significantly easier in the North as compared with the South. As a result, there were far more variations of spoken Chinese in the South. This is what led the new government to choose an already-unified language cluster as the basis for the national dialect. As a result, the larger portion of the population had to change very little, and only the fragmented South (precise definition in § 5.2) needed to adapt.

2.2 Current Chinese Phonology

Due to having thousands of years of history, the phonology of Chinese has undoubtedly changed in many ways. The scope of this thesis is not concerned with ancient Chinese, so only the standard pronunciation, *Pŭtōnghuà* will be discussed. Many languages have a large overarching encyclopedia of phonology, and for *Pŭtōnghuà*, this work was created by Duanmu (2000) and now serves as the seminal work for phonological research in Standard Chinese. The concept of tonality will be largely ignored, as it will be unused for this thesis.

2.3 Syllable Structure and *Pīnyīn*

Chinese syllables consist only of an initial and a final. Initials can be any consonant sound or a vowel. The finals can only be vowels or a limited number of nasal sounds. In other words, the only necessary component to a Chinese syllable is a vowel, such as in words like *è* 饿 *(hungry)* or *ài* 爱 *(love)*. Additionally, the phonotactics of Chinese do not allow for any possible consonant clusters as they exist in English, so the *Pīnyīn* digraphs above still constitute only a single sound. Because the only consonants Chinese allows as finals are [n] and [ŋ], that means there are mostly vowel final sounds. To dissect vowel finals more accurately, they can be described with the rule, where items parenthesized are optional.

$$\text{finals}_{\text{rimes}} = \text{(medial)} + \text{vowel} + \text{(terminal)}$$

(Sun 2006: 37)

For the common particle ending *liăo* 了(*finish*), the [i] sound is considered the on-glide, the [ɑ] is the main vowel, and the [u] is the off-glide. The best rule of thumb for deconstruction of diphthongs and triphthongs in Chinese is to look at the *Pīnyīn* diacritic over the letters: whichever mark has the diacritic is considered the main vowel.

This thesis will use the widely accepted romanization scheme that is used all over the world, *Hànyŭ Pīnyīn*, usually shortened to *Pīnyīn*. This system was created in the 1950s following the dilemma of constructing a national language that was described earlier. Consisting of only Latin letters, it is how non-native speakers learn how to "standardly" pronounce Chinese sounds.

Pinyin	IPA	Pinyin	IPA	Pinyin	IPA	Pinyin	IPA	Pinyin	IPA
b	[p]	q	[tɕ^{h}]	i	[i]	a	[a]	ui	[wei̯]
p	[p^{h}]	x	[ɕ]	in	[in]	ai	[ai]	ua	[wɑ]
m	[m]	z	[t͡s]	ing	[iŋ]	ao	[ɑu]	uai	[wai̯]
f	[f]	c	[t͡s^{h}]	u	[u]	an	[an]	üan	[ɥɛn]
d	[t]	s	[s]	ong	[oŋ]	ang	[ɑŋ]	uang	[wɑŋ]
t	[t^{h}]	zh	[t͡ʂ]	iong	[joŋ]	ie	[iɛ]	üe	[ɥɛ]
n	[n]	ch	[t͡ʂ^{h}]	ü	[y]	iu	[jou̯]		
l	[l]	sh	[ʂ]	e	[ɤ]	ia	[ja]		
g	[k]	r	[ʐ ~ ɻ]	ei	[ei̯]	iao	[jɑu]		
k	[k^{h}]	y	[j ~ ɥ]	ou	[ou̯]	ian	[jɛn]		
h	[h]	w	[w]	en	[ən]	iang	[iɑŋ]		
j	[tɕ]			eng	[əŋ]	uo	[wo]		

Table 2.1: Transliterations of initials (first two columns) and finals (last three columns) in *Hànyŭ Pīnyīn* alongside IPA

2.4 Glides and Approximants in the Thesis

An important discussion that is relevant for this thesis is when exactly the sound [w] becomes present in Chinese. With the information already described, Chinese syllables can consist of only a vowel, only a diphthong, or these sounds can be present at the beginning of the syllable with a limited number of nasal finals. The sound [w] is phonetically very similar to the vowel [u], as will be described later, so this sound is only present when there is an initial [u] sound taking the place of the on-glide. For example, the word *wáng* 王 *(king)* can technically be described phonetically as [uaŋ] and the word *wài* 外 *(outside)* can be written phonetically as [uaɪ]. Of course, both transcriptions do not take into consideration any stress, or durational variation in the on-offglides. In terms of *Pīnyīn*, the /w/ is present at any point that there is no initial, and the medial is [u] (Sun 2006). In general, a good definition of what a glide is that this thesis follows can be provided by Urua & Udoh (2017) in which they describe them as transitional sounds which veer either toward or away from another sound; the nature of glides is loosely like consonants due to the lack of steady states. This definition supports the usage of describing [w] in this thesis as a glide.

In his seminal work on Chinese phonology, Duanmu (2000) does not directly give any acoustic analysis for this sound, but he does mention the presence of the sound change from [w] to [ʋ] and some of the features that accompany [ʋ] after the sound change has occurred. He primarily associates the sound change with the Beijing region, which will be discussed in § 5.2. In his chapter on the sound inventory, he elaborates on the possible onset and rime combinations for both [w] and [ʋ] as well as why some combinations are much more likely to undergo this sound change than others are. For example, the word *wǒ* 我 (*I/me*) will almost never undergo a sound change to [ʋǒ] because both "[w] and [o] share the feature [+round]" (Duanmu 2000: 23). This line of reasoning can further be extrapolated to the [wu] onset rime combination, as in the word *wǔ* 五 *(five)*. Lastly, he suggests that this sound change is prevalent not only in pure Chinese vocabulary but can permeate into English when spoken as a second language. The precise acoustic features of this sound change as well as the sociolinguistic characteristics that this thesis is based on will be deeply analyzed in chapters 4 and 5, respectively.

3 General Acoustics

This section will focus on the scientific and mathematical basis behind theories of acoustic phonetics. I have organized this section by first discussing the fundamentals of wave physics, then the translation of pure sounds into spectrograms, then a description of source-filter theory. This section concludes with some information on how these acoustic features manifest themselves in certain phones, which is the focal point of Chapter 4.

3.1 Fundamentals of Wave Physics

Sound can be colloquially described as wiggly air. What creates sound waves is variation in air pressure in relation to some source. Sound waves are emitted as longitudinal waves from whatever the source is, meaning that the waves travel in the same or opposite direction as the air molecules themselves. These waves can be translated into a transverse wave, meaning that the movement of the air particles is perpendicular to the direction that the wave travels (Reetz & Jongman 2009). Because we cannot see the individual air particles during speech production, a translation mechanism from the physics is necessary. A depiction of this translation between longitudinal and transverse waves is shown in Fig 3.1. The darker vertical bands are moments where the air pressure is higher, and as the lever repeatedly pushes air through the apparatus, locations of higher and lower air pressure are created. The resultant wave that most people are more familiar with maps air pressure variation on the vertical axis. Another word for the mapping of these transverse waves is an *oscillogram*, which quickly become very complicated, given a complex sound.

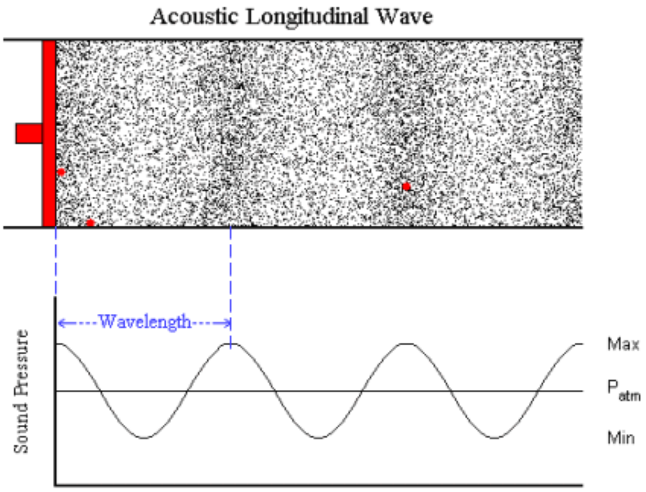


Fig. 3.1: Longitudinal wave (top) and its corresponding transverse wave (bottom) (Source: Institute of Sound and Vibration Research, University of Southampton)

This depiction of sound waves is an important step in understanding deeper complexities of sound propagation, and later, source filtering. Once the transverse wave has been translated, defining some important terminology can now get underway. The *wavelength* (λ) of a sound wave is the distance from two analogous portions of the transverse wave. The location could be the 0 crossing, a peak, a trough, or some other location along the curve, so long as the two points remain 360° apart. Wavelength is measured in distance, usually meters.

The *period* (T) is the amount of time that a wave takes to make a full 360° cycle to return to its original position. The period of a sound wave is directly related to the wavelength, as the two are directly proportional; as one increases, the other will also increase. Lastly, *frequency* in basic acoustic terms is more complicated than the previous two and requires a dependency unit of measurement. A sound wave's frequency is how many periods that the wave repeats itself, given a certain standard of time, usually 1s (Equation Set 1). In speech sciences, the frequency is the most common measurement, as it is what will later be used to measure things such as pitch and formants.

$$Frequency\ (f) = \frac{number\ of\ periods\ (T)}{1\ (s)}$$

Equation Set 1: Definition of wave frequency with respect to periods

All these representations of period, wavelength, and frequency are inter-connected. As can be seen from the Equation Set 2, the relationship between them all is very important, and calculations for one of them can be extrapolated to calculate some part of the others (Reetz & Jongman 2009). In the following equations regarding sound wave physics, c represents the speed of sound ≈ 343 m/s, while in other applications of particle physics, c instead represents the speed of *light*.

$$\lambda = \frac{c}{f} = c \times T = c \times \frac{1}{f}$$
$$\frac{c}{\lambda} = f$$
$$\lambda \times f = c$$

Equation Set 2: Relationships between wavelength, frequency, period, and speed of sound waves. c ≈ 343 m/s

Next, knowing how to read an oscillogram is important, but not every element of speech can be detectable with just the waveform, such as voice onset time and noise. This thesis primarily utilizes the spectrogram to obtain its data, so a brief explanation of how a spectrogram is created from an oscillogram is more important. Real speech signals appear as a jumbled mess of repeated waves that are impossible to decode without the aid of a computer. A waveform must undergo a complicated sequence of events called a Fourier analysis to deconstruct all the various frequencies within a waveform into separate parts visible on one graph called a spectrum. A pure sine wave would be deconstructed into a spectrum with only a single point, as can be seen in Fig. 3.2. However, when there is a composite function created by multiple sine functions, the wave becomes more distorted and other frequencies begin to appear on a spectrum, as shown in Fig. 3.3.

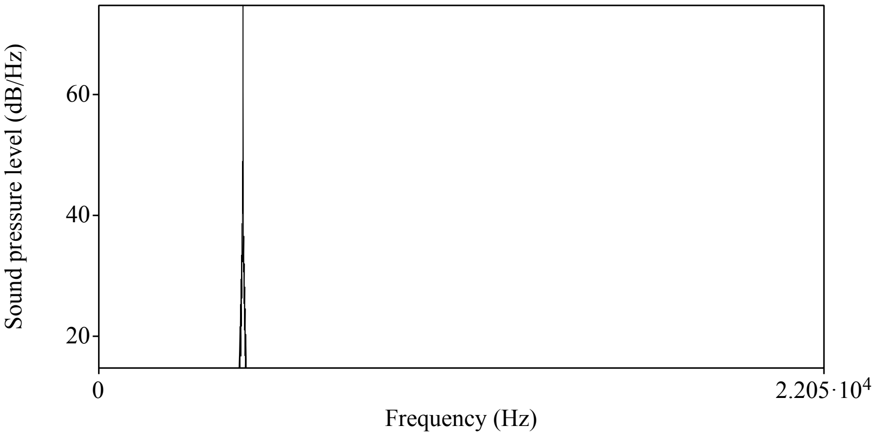


Fig. 3.2: Pure sine wave's spectrum at 4400 Hz

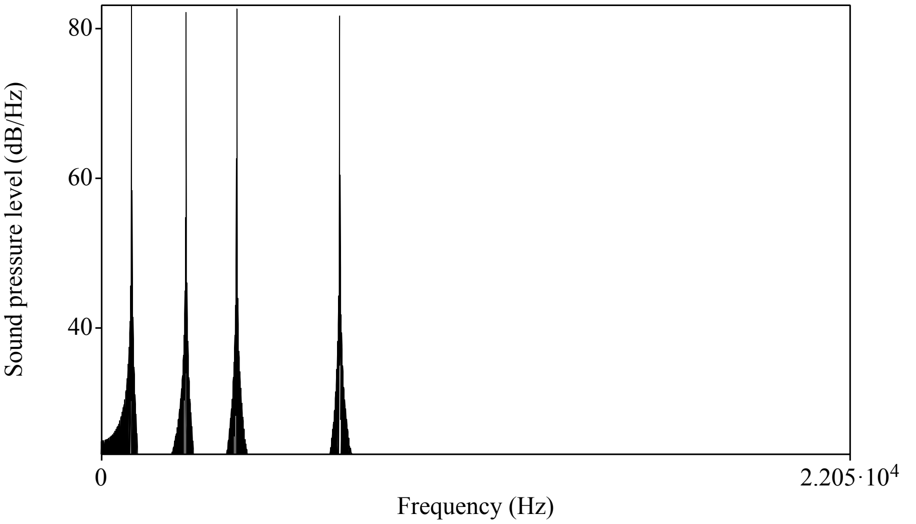


Fig. 3.3: Composite spectrum from four pure sine waves with different frequencies

Spectrograms are 3-dimensional; to create one, turn the spectrum onto its side and extend along the z – axis for the duration of the sound. For a pure sound or a composite mixture of pure sounds, the spectrogram will show only long dark horizontal bars representing the frequencies found in this sound, as shown in Fig. 3.4. In this way, using a spectrogram to analyze pure tones is no better than using a power spectrum. However, a spectrogram becomes much more useful in real speech, as our vocal folds do not create pure sounds, and are further filtered through our vocal tract, as § 3.2 will describe.

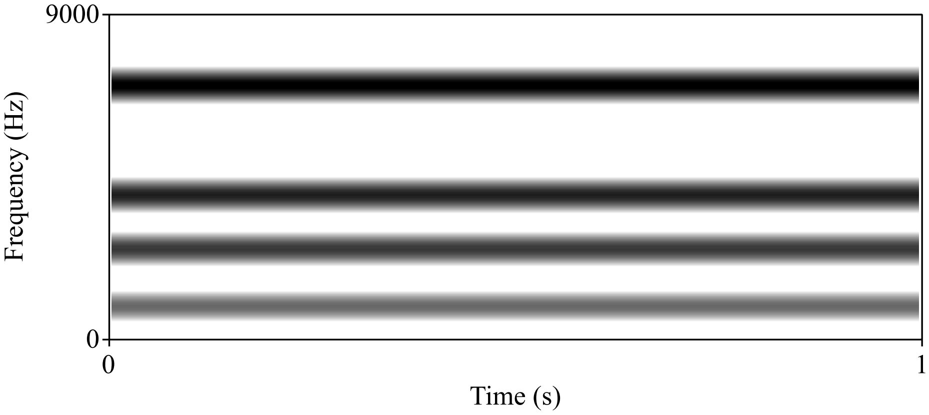


Fig. 3.4: Spectrogram of the composite sine waves shown in Fig. 3.3

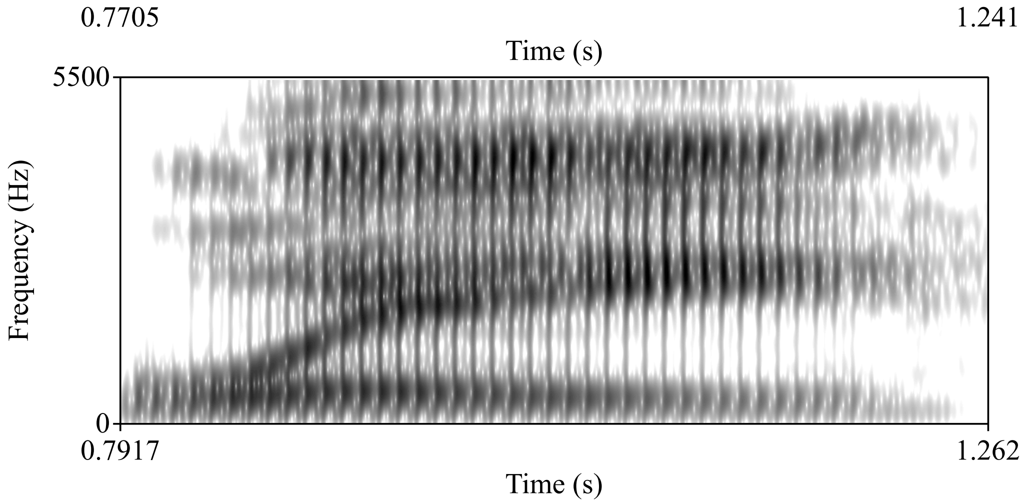


Fig. 3.5: Spectrogram of American Male saying the word *way*. Notice the steep movement of F2.

3.2 Source-Filter Theory of Speech Production

As alluded to in the previous section, humans are unable to produce pure sine waves as sound. Instead, the vibrations produced by our vocal folds are augmented on their way out of the speaker by the shape of our vocal tract. In other words, the vocal folds are the *source* of a sound (with its own inherent variations) and the length of the vocal tract above the glottis is the *filter* that shapes the sound into something that sounds like a human (Reetz & Jongman 2009). This very well-established pattern of sound production was created by Fant (1960) using Chiba & Kajiyama's (1941) seminal work on X-ray structures of the vocal tract. Though technically considered a *theory*, Source-Filter is one of the most prominent models used to describe the production of human speech. The variability of the filter is caused by the natural variability of the vocal tracts of humans. Fig. 3.6 is a depiction of the source-filter theory framework, showing an unfiltered spectrum, a filter with the shape of formants, and the resultant output power spectrum.

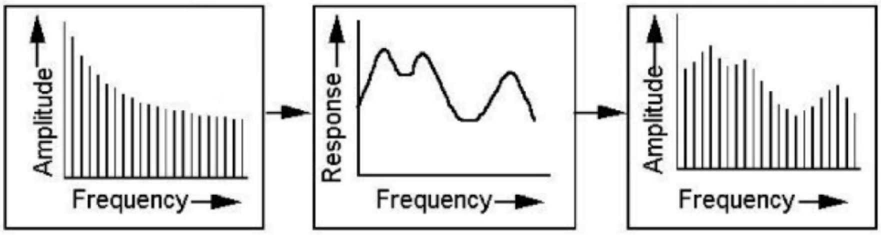


Fig. 3.6: Frames from left to right are the source sound, the filter, and the output sound signal (Ballew 2019).

3.2.1 Tube Resonators

Understanding how a source and a filter work together is important, but even more imperative to understanding this thesis is how tube resonators function and how that is representative of what sounds humans can actually produce. A commonly used descriptor of how the vocal tract is like a set of cylinders is called the *Twin-Tube Resonator* theory. This theory suggests that the vocal tract is separated into two differently sized cylindrical tubes, perhaps of varying lengths, with the connection between them also varying in size and position. This is meant to represent the different tongue positions within the mouth changing the exact constriction location for articulation of various vowels (Johnson 2012).

Fig. 3.7 is a simplified diagram showing the single constriction within the vocal tract during a vowel articulation such as [a]. The left side of the graph, l_b, represents the bottom portion of the vocal tract, i.e., the length of the esophagus from the vocal folds up to the back of the mouth before the tube greatly widens. The right side of the image, l_f, represents the length from the back of the velum to the exit of the mouth. The calculation for the resonance of such a tube can be calculated with the following formula,

$$F_n = \frac{c(2n-1)}{4L}$$

Equation Set 3: Calculation of formant resonances

where c is the speed of sound, n is the number of resonances that can fit within the tube area, mandatory to be an odd number for open cylindrical tubes, and L is the length of the corresponding tubal section. Any math beyond this one formula is unnecessary for this thesis, but an in-depth mathematical basis for this type of sound propagation through cylindrical tubes can be found in chapter 3 of Stevens (1999).

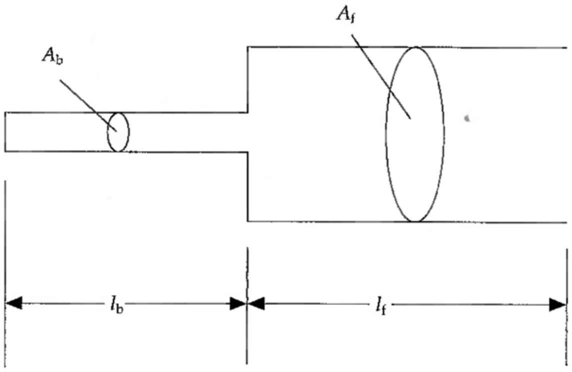


Fig. 3.7: Two-tube resonator model of the vowel [a]. A_b and A_f represent the cross-sectional area of each portion of the back or front cavity resonators.

The case of resonance for the sound change of this thesis specifically will be discussed in § 4.3. Fig. 3.8 is an image of possible resonances of a typical human vocal tract, suggesting that tube lengths of both sides vary with different articulatory locations, which thus leads to different output formants. These are twin-tube models of various articulations of vowels and the resulting formant patterns that arise because of constrictions in different locations. These constrictions are represented by either the narrowing or widening of the pipes and are analogous to the tongue moving to different positions to change the flow of air. However, an important note is that formant frequencies are built-in to the vocal tract of a speaker; tongue placement can affect which ones are filtered, but the formants of a speaker do not change with variations in amplitude or pitch.

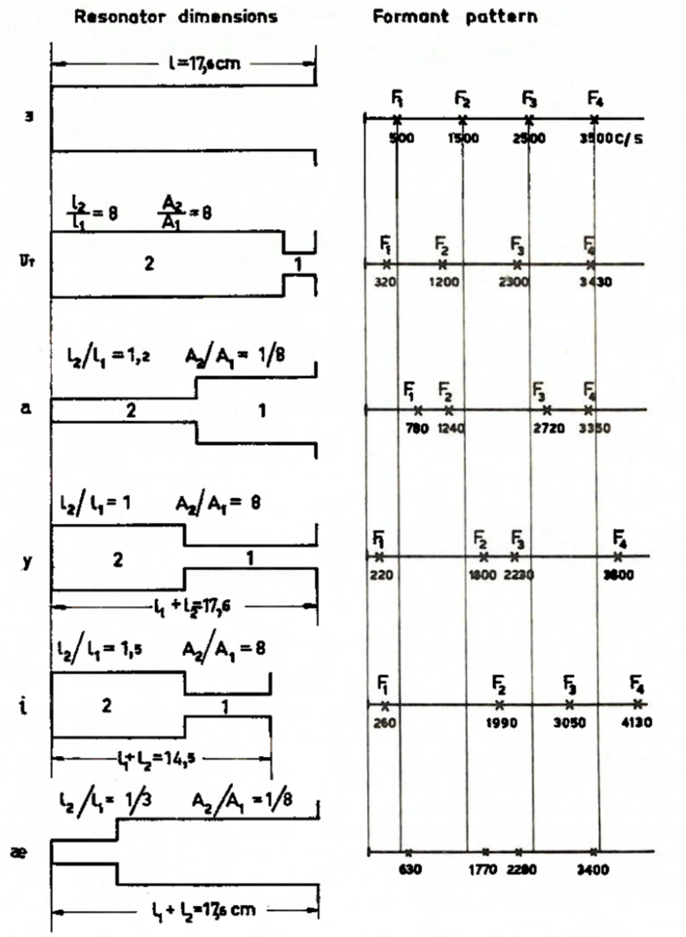


Fig. 3.8: Different resonance constriction locations and how they are an estimate for formant frequencies (Fant, 1960: 66).

The ways in which our vocal tracts are manipulated influence the resulting acoustic parameters that can be measured. Thus far, there has been much theoretical build-up to ensure that it is easily understood what is happening during the production of certain sounds. This thesis is only interested in two sounds, namely [w] and [ʋ] in *Pǔtōnghuà*; the next chapter will do a detailed analysis on the phonetic and phonological characteristics of these sounds.

3.3 Harmonics to Noise Ratio

An additional useful feature of speech analysis used here is the harmonics-to-noise ratio. HNR is fully defined as “the relationship between harmonic and noise components” of a sound segment (Fernandes et al., 2018) and more briefly by Boersma (1993) as a sound’s degree of periodicity. In practice, HNR is a parameter that measures what percentage of a sound’s energy is caused by a voicing, thus producing harmonics, and how much is just noise. HNR is a good indicator of voice quality and has seen applicable uses in things such as pathologies of some vocal tract injuries and disease (Teixeira & Fernandes 2015).

The mathematics for calculating HNR can be found in § 8.3. A good depiction of what is considered noise can be seen in voiceless sounds. The image below shows the waveform of a clear fricative with no apparent repetition of waveforms; it appears as a jumbled mess of indiscernible signal: this is noise. Most of the energy in this sound signal is therefore present in the aperiodic components. A spectrogram would be equally as informative for showing noise, but a waveform is used here because the second half, simple periodicity, is easier to see in the waveform. The sound in Fig. 3.9 will have a very low (if not negative) HNR.

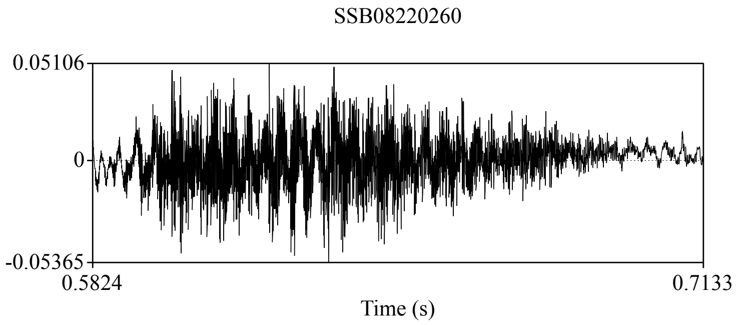


Fig. 3.9: A selected waveform from the affricate [ts^h]

The next waveform is one that is mixed with clear periodic and aperiodic signals. This is meant to show how in any voiced sound, there is some percentage of noise and periodicity. A signal such as this will have a majority of its energy distributed within the repeated signals. The sound in Fig. 3.10 would likely have an HNR above 15 dB.

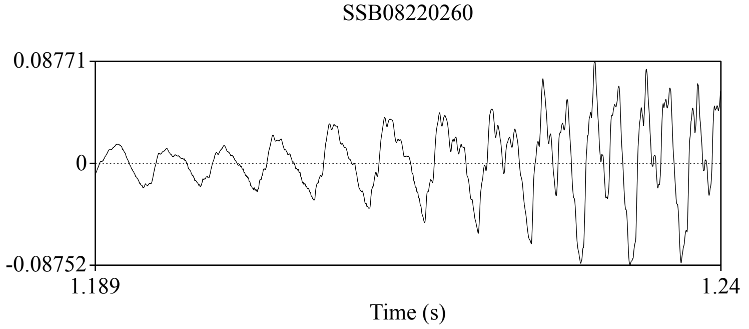


Fig. 3.10: Clear mixture of periodic and aperiodic signals creating this composite sound.

Discussing HNR may be a very valuable tactic in acoustically differentiating between [ʋ] and [w]. The articulation of [w] involves a mostly laminar airflow. The constriction at the velar region and labial region do not impede airflow; they simply redirect it. As a result, the energy remains constant. However, the articulation of [ʋ] involves an additional constriction of the upper teeth and lower lip, which impedes airflow, even if only for a brief moment. That moment creates a break in the periodicity, resulting in turbulent airflow, and therefore more noise. An articulation of [w] will produce an HNR that is significantly higher than that of [ʋ]. This will be one of the key elements analyzed later in this thesis.

4. Approximants and Sound Change

The previous section ended with a description on why studying HNR is important for distinguishing between the two variations of [w] and [ʋ] respectively. In this section, I will discuss some of the features of approximants broadly, their articulations, and how that is mapped onto specific acoustic characteristics of these sounds.

4.1 Scope of Approximants and [w] and [ʋ] Broadly

Despite being a unique set of sounds in terms of acoustics and articulation, approximants are quite common throughout the languages of the world, irrespective of region. The labiovelar approximant [w] is common in many languages, occurring in 75.7% of languages surveyed, but the labiodental approximant [ʋ] is much rarer, with only 1.9% of surveyed languages using this sound (Maddieson & Disner 1984; Colombo 2015). However, an important note is that these numbers only count these sounds when phonemic, and do not investigate sound change at all. Some examples of this sound occurring in languages can be found in Danish, Dutch, and of course, *Pǔtōnghuà*. Danish, for example, describes [w] as being in free variation with [v], but only allophonically with [ʋ] (Basbøll 2015). Additionally, Dutch is a language that uses this sound, but I have found conflicting information on its phonemic status. Booij (1999) simultaneously specifies the presence of [ʋ] within Dutch under *allophonic variation*, but later classifies it as a standard phoneme. However, one interesting usage of [ʋ] in Dutch is that it can occur in word-final position. This is in direct contrast with the phonotactics of *Pǔtōnghuà*.

Pǔtōnghuà only uses the [w] phoneme in syllable-initial position and can only be followed by a specific number of vowel rimes; as a result, [ʋ] only occurs syllable-initially as well. Sound change in general, and then with respect to *Pǔtōnghuà* [w] will be discussed in the following sections. The effects of rime in this sound change are discussed in § 4.4.

4.2 Sound Change

Sound change, sometimes referred to as phonetic change or phonologization (Hyman 2013; Kiparsky 2015), is the replacement of one sound with a different one. This can occur in any language and there have been numerous studies that examined things such as different types of sound change as well as what may be the underlying cause of these sound changes. Garrett & Johnson (2011) describe the phonetic biases that influence why some sound changes occur more than others and mention three key characteristics that any person studying sound change ought to consider: typology (rare vs. common changes), conditioning (lexical-morphological influences), and actuation (phonetic triggers). These authors explain four factors that can broadly influence the prevalence of sound changes from a phonetic basis. First is motor planning, suggesting that some sound changes are simply merged due to likeness of articulators, and are therefore simpler to pronounce. The second consideration is aerodynamics, positing that the air pressure variation spectrum between vowels and stops (Ohala 1983) can influence which sound changes are more likely to occur. The third is dubbed gestural mechanics. This is like motor planning but relies on variation between articulators; distant co-articulators may either merge into a central articulation or one may take over completely. Last is perceptual parsimony, which suggests that sound changes can occur because of ease of perception/hearing.

The sound change under question for this thesis is the change from the labial approximant [w] into the labiodental approximant [ʋ]. The exact reason that this sound change occurs is beyond the scope of this thesis, but some possible sociolinguistic influences will be elaborated on in chapter 5. As mentioned earlier, [ʋ] does exist phonemically in some languages, but primarily arises due to allophonic variation. In the case of the sound change specifically in *Pǔtōnghuà*, the labiovelar variety is considered phonemic while the labiodental variety is considered allophonic (Wiener & Shih 2011). This sound change has been shown to occur in Standard Mandarin (Chao 1927; Jiong 1987; Duanmu 2000; Hu 1991); despite not being standard to *Pǔtōnghuà*, it is an often-recognized pronunciation variant.

4.3 Specifications of the Approximants [w] and [ʋ]

The articulation of [w] is a bit nebulous, primarily due to its requisite coarticulation dependencies from opposite sides of the mouth; labial and velar combinations making proper classification difficult. There were questions whether it should be described as primarily a velar, mostly a labial, or some combination of the two. Many early phonologists would categorize [w] either as only a labial or only as a velar, but Ohala and Lorentz (1977) were one of the first, following Ladefoged (1971) to suggest that [w] ought to be categorized as a coarticulated *labiovelar* for two reasons. First, many languages, particularly those categorized within the Niger-Congo language family, already utilize labiovelar clusters such as /kp/ and /gb/ that are relatively foreign to the commonly studied Indo-European languages. Second, they discuss the acoustics behind pure labials, pure velars, and labiovelars, and prove that these three sound categories produce strikingly similar acoustic characteristics (Ohala & Lorentx 1977). Finally, Recasens Vives (2014: 35) suggests that the standard F2 of the acoustics of [w] is around 1000Hz. However, this does not take into consideration gender or age differences. Additionally, this value may be analogous to this thesis's definition of $F2_{min}$: the 1000 Hz value is likely taken from the stagnant moment of articulation of the [w] before the on-glide, which is closer to the articulation of [u].

This description of [w] coarticulation gains further merit when the classification scheme for other sounds is analyzed. The way to describe the place of articulation of most sounds is to consider which location is the most radical or prominent component of said articulation. This generally works well for most articulations, such as [t] being a clear alveolar and [k] being a clear velar, but quickly falls apart when describing approximants such as [w] and [ʋ] (Anderson 1976). Considering [w] to have equal articulators makes more sense when the acoustics of this articulation are discussed.

4.3.1 Correlation Between Articulation and Acoustics

Certain articulations directly map onto certain acoustic patterns. Possible formant patterns arise as a function of a person's vocal tract, and the articulation and manipulations of sound waves within the vocal tract highlight different sections of those possible resonances (Fant 1960). In other words, different tongue positions will change the acoustic output that can be measured via our technology. This section will discuss the acoustic patterns that arise with the articulation of approximants and more specifically the labiovelar and labiodental articulations.

Approximants are known to have a few key features, with the labiovelar approximant [w] having some of its own. The primary characteristic of the labiovelar approximant is its tendency to lower all formants. Stevens (1998) found that the combination of lip rounding and a raised tongue body in the velar region together attribute to a lowering of all formants, but especially F1 and F2. The lip rounding increases the length of the vocal tract, allowing for a slightly deeper resonance, and the velar pseudo-articulation creates a secondary constriction near the middle of the vocal tract, as shown in Fig. 4.1. The mathematical basis behind these resonances can be studied in more details in Stevens (1998) and Fant (1960).

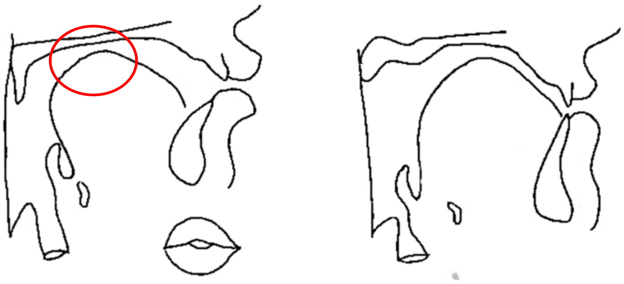


Fig. 4.1: Depiction of a [w] articulation (left) and a [j] articulation (right) to show the semi-constriction of the vocal tract into two separate tubes (Stevens 1998: 516)

Additionally, Reetz & Jongman (2009) provide a spectrogram example of what this formant lowering capability physically looks like. Fig. 4.2 below shows a spectrogram of a male saying the phrase [iwi].

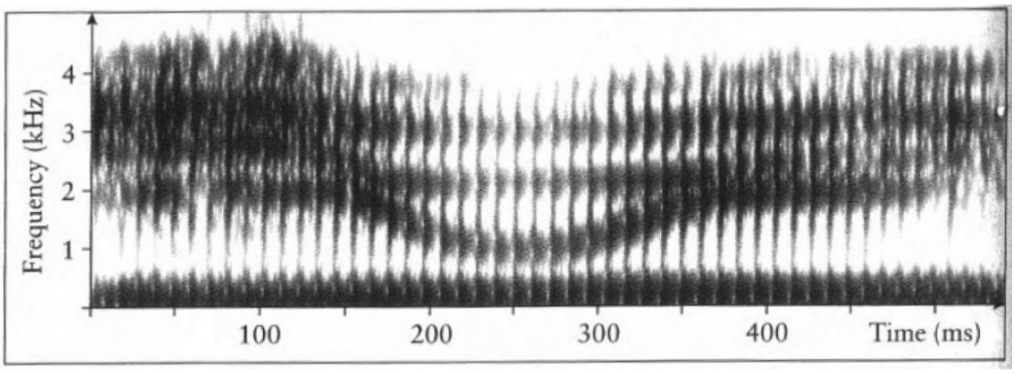


Fig. 4.2: Adult male speaking [iwi]

This characteristic of [w] also translates to the variant [ʋ] due to sound change, but the formant lowering capabilities of this sound are dramatically reduced, as the lip protrusion no longer exists and there is no longer a prominent velar component, as the two primary articulators are now the lips and teeth. This is the key difference between [w] and [ʋ] that is used to analyze the data of this thesis. Assuming that other variables are adequately controlled for, an utterance beginning with [ʋ] will produce (1) an F2 value that begins at a much higher frequency than the [w] variant and (2) an F2 slope that is significantly flatter than the [w] variant. As shown in Fig. 4.3, the difference between a prototypical [w] articulation and a prototypical [ʋ] articulation is very prominent in the spectrogram. This is the theoretical basis of this thesis.

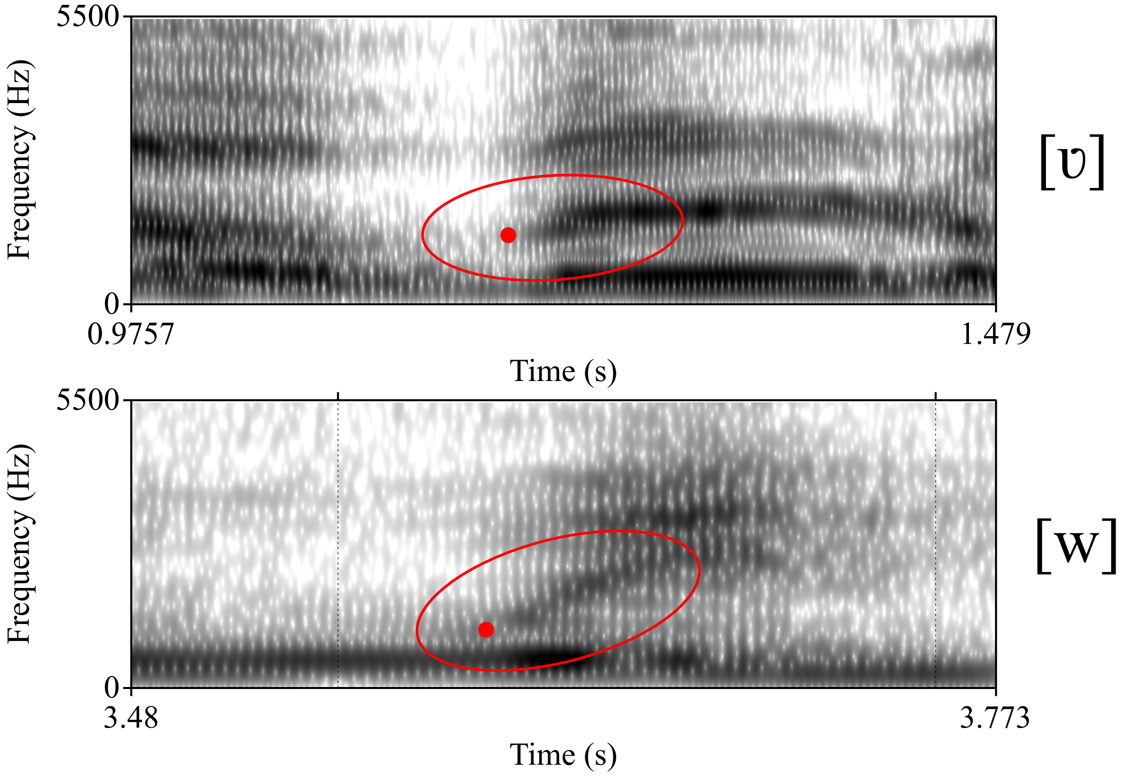


Fig. 4.3: Articulation of the Chinese word wèn 问 (*ask*) articulated with [ʋ] (top) as well as with [w] (bottom). Notice the prominent glide for [w] and a lack of slope for the [ʋ] variant as well as a slightly lower starting location of F2 for [w].

4.4 Onsets and Rimes

§ 2.4 briefly explained onset and rime patterns in *Pǔtōnghuà.* This is the organizational scheme that dictates the prevalence of this sound change in this language, which are allowed or not allowed by Chinese phonotactics. Chao (1927) laid the groundwork for onset and rime combinations in Chinese, and his work has been elaborated on many times later. One of the key determinations of his work and others like Shen (1987) and Hu (1991) is that there is a very clear delineation of which onset + rime combinations produce this sound change the most. It has been found that the combinations of [w] + /ei/ and [w] + /en/ produce the most labiodental variant [ʋ] among all possible combinations. Wiener & Shih (2011) repeated many of these measurements to a similar result and further discovered that [w] + /en/ produced the most, while [w] + /ei/ produced the widest variation in pronunciations. For these reasons, these are the two onset + rime combinations that this thesis studies. Later, these two combinations will just be referred to as the words /wei/ and /wen/, irrespective of tone or semantic meaning when translated into English. Lehiste & Peterson (1961) produced data indirectly corroborating a similar pattern in English. Though not the key premise of their study on diphthongs, they showed that even in English, [w] + /ei/ as in the word "way" produced much variability in pronunciation.

5. Demographics and Sociolinguistic Considerations

As mentioned multiple times, this thesis performs its acoustic analysis around two demographics (and a third predictor), namely sex, region, and age. The sound change from [w] to [ʋ] in *Pǔtōnghuà* has been established as existing mostly within the demographic of young females in Northern regions of China (Hu 1987; Shen 1987), so this chapter will discuss each of these demographics independently, what features are being taken into consideration, and any possible complications involving these demographics.

5.1 Sex and Gender

The terms *gender* and *sex* are both not the most accurate term to use for this thesis. Gender is the term often used for research of this type, but when studying something physical about a human, such as aspects of speech that are a direct result of their anatomy, a sociologically based term does not check any of the appropriate boxes. However, the term *sex* itself is not a binary, even biologically and genetically, so there are also issues in labelling the prototypical male/female divide with this term. This thesis will utilize the word *gender* to abide by standard conventions within linguistics research, with the acknowledgment that it does not accurately describe the male/female divide as accurately as once thought. This thesis is sociophonetic in nature, so an explanation of the sociological differences between these terms is still worth discussing.

The most prominent difference between the genders with respect to speech analysis is the lowering of a fundamental frequencies in conjunction with a lengthening and thickening of vocal folds (Zamponi, V. et al. 2021). Males typically have larger vocal cords, and in conjunction with a longer vocal tract length, produce formant frequencies that fall lower than the range of the average female speakers. For this reason, it is expected that the data presented in the results section for minimum F2 frequency will trend lower for male speakers and higher for female speakers. However, aside from simple acoustic variation between males and females, there have been previous studies that social stratification across gender lines can influence pronunciations of standard words (Graddol & Swann 1989: 61). These two authors write that not only is pronunciation variable across gender lines, there may also be complementary interactions that affect pronunciation, such as region and age, all of which are considerations that are present in this thesis. Despite Graddol & Swann's (1989) study being about English, the trends can be extrapolated to other languages.

The notion of naturally higher formant frequencies in females is only exacerbated within the confines of this sound change. Multiple researchers in the past have shown that when using *Pǔtōnghuà*, women are significantly more likely to undergo the transition from the labiovelar [w] to the labiodental [ʋ] than are men (Shen 1987; Hu 1987). This increased usage of a sound that naturally produces higher frequencies further creates a divide between males and females. However, the difference between genders for this sound change may not be quite as prolific as previously suggested by Shen (1987) and Hu (1987). Zhou (2003) performed a detailed analysis on the speech of news casters across China and found that women used [ʋ] in 68% of analyzed segments and men used [ʋ] 64% of the time. It should be noted that this does only take into consideration news anchors in China, which almost always fall under very standard speakers who are both young and highly educated. In other words, there is not much variation. This is one of the only studies that posits the presence of [ʋ] in males.

In another study of his, Hu (1991) summarizes some of the features that make speech sound feminine in China. He described a speech that sounds *fragile* as one that is more representative of Chinese femininity; within his participants, not a single man used articulation patterns indicative of his definition of feminine. Within the scope of that paper, he describes fronted and smaller articulations as being considered fragile and feminine, which tightly corroborates Chao et al.'s translated notes (1937), which suggests *nǚguóyīn*, 女国音 (*feminine accents*) to be a result of the dentalization of the alveolar and palatal sibilant series, so much so, that they considered it to "be a gendered pronunciation." Socially, the gender divide within China is quite prominent, and any attempt at being perceived as more feminine (or more masculine) is often undertaken, even if in a subconscious, unintentional way (Hu 1991: 53). This driving force for feminine pronunciation is further corroborated by additional studies, such as those that directly tie specific articulatory features as being indicative of femininity, discussed next.

Hu (1991) described it as being polite for women to speak and laugh with a small mouth, which limits the possibility of producing more protruded lips and a deeper sound. Additionally, there has been a suggestion that physical appearance may further increase the usage of [ʋ], as "visual appeal may [be] an additional factor insofar as there is less lip movement for [ʋ], thus yielding a more demure and reserved countenance" (Chan & Lin 2019). The exact articulation features were described in an earlier study by one of the same authors. Chan (1996a) explains how the [ʋ] variant utilizes spread lips and that the teeth being closer to the lower lip causes a flatter, less rounded articulation than the standard [w]. She reiterates what has been established in a previous section about standard [w] producing a lower acoustic frequency than [ʋ] because of lip rounding.

In conclusion, the juxtaposition of these sociological and biological characteristics creates the ideal circumstance with respect to the sound change that this thesis is investigating. Because of these combinations of features, women, at least in most cases, are thought to undergo this sound change more frequently than men. This is due to the fragile perception of femininity as well as the robust articulatory traits that accompany those representations of Chinese femininity. Based on this section, it is apparent why this thesis is incorporating gender differences to such a heavy extent in the data analysis and interpretation; it is a very important feature in understanding the extent to which this sound change has permeated outside of the typically studied young, northern female demographic (Hu 1991).

5.2 Region

The division of North and South within the borders of China is an important distinction. While the corpus used for this thesis does not explicitly define what is meant by their delineation between North and South, it can be assumed that a relatively standard distinction was made. It is very common for the latitudinal separation of China to fall on the Qinling-Huaihe line. This is the sociodemographic line nationally recognized since 1958 as the demarcation that separates Northern and Southern China (Liu 2020). The line runs from the Qin Mountains in the center of China to the Huai River that runs just north of Shanghai. Fig. 5.1 shows a map of this line. The Huai River off-branches slightly from the Yangtze River North of Shanghai. It is safe to assume that something described as North in this paper refers to something North of this line and vice versa for South.

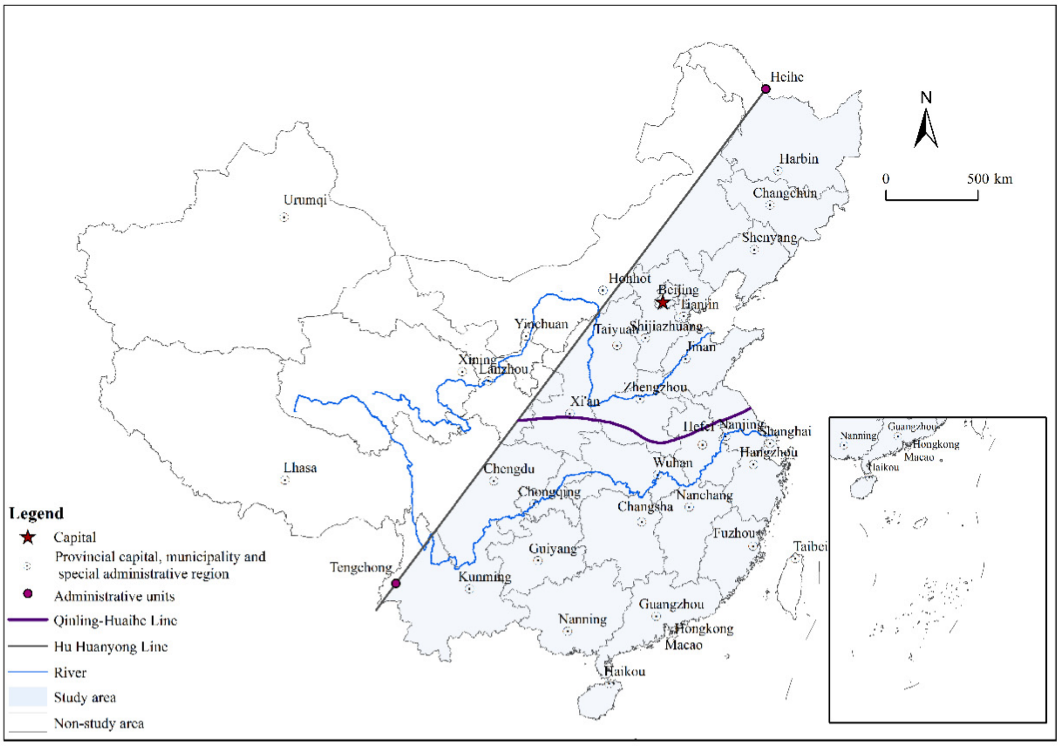


Fig. 5.1 Map showing the Qinling-Huaihe line in addition to other geographic locations in China (Liu 2020).

As was described in § 2.1.1, the phonological basis of *Pǔtōnghuà* was set to be the Beijing variety of speech, which falls in the North (Sun 2006). Because of this, it may be easy to assume that the standard pronunciation baseline of a country is not susceptible to phonetic changes over time. However, this is not the case, as previous literature has pointed out that this sound change exists primarily in Northern regions of China (Hu 1998; Hu 1991; Wiener & Shih 2011). Another example of a region farther north that is also susceptible to this sound change is the city of Harbin. Wang (2011) studied the social factors that are responsible for this sound change in the city of Harbin, which is a city significantly farther North and discovered that speakers there were just as susceptible to undergo this sound change as the speakers based in Beijing. A case may be made for a more widespread sound change that is not limited to the far north region of China.

First, Zhou's (2003) phonetic variation paper reported that Chinese news broadcasters used a labiodental variation in their speech on w-initial words. There was only a distinction here between gender, not region. In other words, she found that there was widespread replacement of [w] with /v/ (her paper did not explicitly use [ʋ]) and is therefore not only isolated to the northern regions. A similar trend was discovered in multiple experimental analyses since Zhou (2003). Additionally, Wiener & Shih (2011) replicated Shen's (1987) work on the Beijing vernacular and found that there was still more usage of [ʋ] in Northern regions, but that the sound change was at least noticeable in Southern regions. Finally, Chen (2010) replicated the word-level analysis of Shen (1987) with an added layer of analysis: region. Chen (2010) incorporated Northern, Southern, and Central regions of China, and found that Southern speakers undergo this sound change approximately 30% less than their Northern counterparts.

An additional factor that ought to be considered is the suggestion that this sound change is not simply a mispronunciation from the standard, school-taught pronunciation. While it is true that the labiodental approximant [ʋ] (or even the non-approximant fricative [v]) does not appear in standard phonology of *Pǔtōnghuà*, it has appeared often enough to be considered a legitimate variation of standard speech. The hypothesis that this sound change arises merely because of mispronunciation in the North is an idea that has been refuted multiple times over by various authors (Zhou 2003; Ying 2011; Lin 1982).

In conclusion, this section has explained why such an emphasis is being placed on regional variation in this thesis. The premise of the current study is to compare the usage of the labiodental approximant [ʋ] for southern regions to something that has been widely established to exist (the North). This section serves as justification for investigating the presence of [ʋ] in non-North regions.

5.3 Age

This section will be brief, as there were not enough speakers in different age brackets in the AIShell Corpus to effectively test the effect of age on this sound change. However, it is an important demographic point that ought to be considered for future research.

The labiodental sound change was historically found in younger generations, but this was something that had been noted since at least Shen (1987), perhaps even as far back as Chao (1927), although his motives and social influences were not at all described. This raises a couple interesting questions. Would the younger generation of the 1980s still use this sound change today, becoming an older generation using this sound? Is this a phonetic feature of younger speakers that people eliminate from their speech over time as they age? These are important questions to truly understanding the depth of where this sound change occurs within China and ought to be answered by other researchers in the future. Lastly, there may be some merit to the idea that this sound change is not limited to China, as a very similar phenomenon has been shown to exist in some varieties of Southern British speakers, all of whom also belonged to the youngest demographic tested (Villafaña-Dalcher et al. 2008). Though this sound did not arise from the labiovelar approximant [w], it did arise from a sound with similar lip rounding and tongue position: bunched /r/. This suggests that a sound change toward the labiodental approximant [ʋ] can arise within the confines of vastly different phonotactics and phonological environments and that young people may be more susceptible to this type of sound change.

6. Methodology

This chapter breaks down all the relevant methods and components of this thesis. It begins with a description of the corpus used, then how participant and sound files were selected before ending with discussing the parameters of how the Praat scripts were annotated.

6.1 The Corpus

The corpus *AIshell 3* offered by *Beijing Shell Technology* (Shi et al. 2020), downloaded from openSLR, contains roughly 85 hours of emotion-neutral recordings by 218 native Chinese speakers, totaling 88,053 individual utterances. This corpus was chosen due to its delineation of data that was conducive to the research question for this thesis. This corpus contained utterances from four age groups (A < 14 years, B 14 – 25 years, C 26 – 40 years, D > 41 years), two region groups (Northern, Southern), and 2 gender groups (male, female). The total breakdown of speakers and to which sub-demographic they belong is shown in Table 6.1.

	NORTHERN	**SOUTHERN**
MALE	31	11
FEMALE	134	40

Table 6.1: Demographic breakdown of 216 speakers in the corpus. Two speakers were removed from consideration because their regional predictor was marked as "other" in the corpus.

Because this corpus was originally created for artificial intelligence research, it contains different amounts of data distributed between both a *test* and *train* folder. The quality and content of the recordings in both datasets is identical, so there was no difference in which set the sound files are chosen from. For the sake of being systematic, sound files were pulled almost entirely from the *train* dataset, with any extra sound file needs being taken from the slightly smaller *test* dataset. This did not have any effect on the results.

Recordings were made with a high-fidelity microphone in a quiet room; the quality is very clear with no background noise. Every sound file contains a single utterance by a single speaker, so every participant has a variable number of sound files associated with them; the recordings are of natural sounding speech of varying lengths. More specific regional information is not available within the corpus, so the only possible classification is *Northern* and *Southern*, which is sufficient for the current study. Every sound file has been transcribed in both Chinese characters and *Pīnyīn* romanization and concatenated into one large comprehensive text file, irrespective of any demographic group. There is one content file for *train* and one for *test*.

Along with a participant and content breakdown, this corpus has a frequency distribution of all tokens. Many of these tokens and words are irrelevant to this thesis, but those that are important have already been described previously as /wei/ and /wen/. The total number of instances of these words appearing in the corpus is shown in Table 6.2. Tone is not studied in this thesis, but they are shown here just to indicate the large number of possibilities for each word.

	TONE 1	TONE 2	TONE 3	TONE 4	TONE 5
WEI	1055	3649	895	4153	8
WEN	762	1137	283	894	4

Table 6.2: Distribution of tones for various words with segmental pronunciations /wei/ and /wen/. Glosses are not possible, as there may be hundreds of different translations of these words.

6.2 Sound Distribution

Due to the large quantity of sound files, it was important to select only recordings that contained the two target words. The specific reason for selecting this sound in isolation was provided § 4.3.1. The entire transcripts of both the *test* and *train* datasets were copied into an Excel spreadsheet. The first column contained only the name of the sound file, and the second column contained the content of that sound file. The second column was filtered to show only those utterances that contained any usage of either /wei/ and/or /wen/. At this point, there is no distinction between gender or region, and only utterances that contain either of the test words remains. In total, there were 9,015 sound files that contained /wei/ and 2849 sound files that contained /wen/. These numbers do not match the totals from Table 6.2 because some sound files will naturally contain multiple instances of the target words.

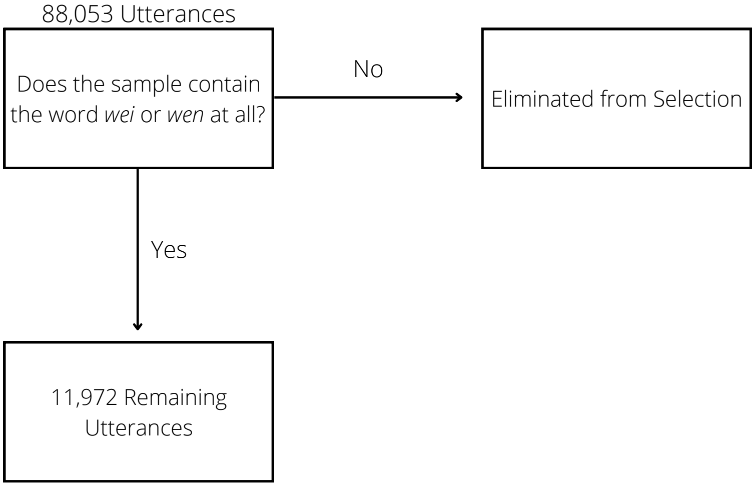

88,053 Utterances
Does the sample contain the word *wei* or *wen* at all?
No
Eliminated from Selection
Yes
11,972 Remaining Utterances

Fig. 6.1: Flow chart showing initial filtering of utterances

6.3 Participant and File Selection

Following the filtering of which utterances contained viable test words, speaker selection was next. To facilitate this, the contents of a provided speaker key (i.e. which speakers belonged to which demographic) were copied into a separate Excel sheet, then sorted by age, gender, and region. The next step was to choose which of the 218 speakers were in the correct age bracket (B), then subsequently organize all the speakers in that age bracket into the four demographic combinations, as in Table 6.3. Figs. 6.2 and 6.3 show this secondary filtering process.

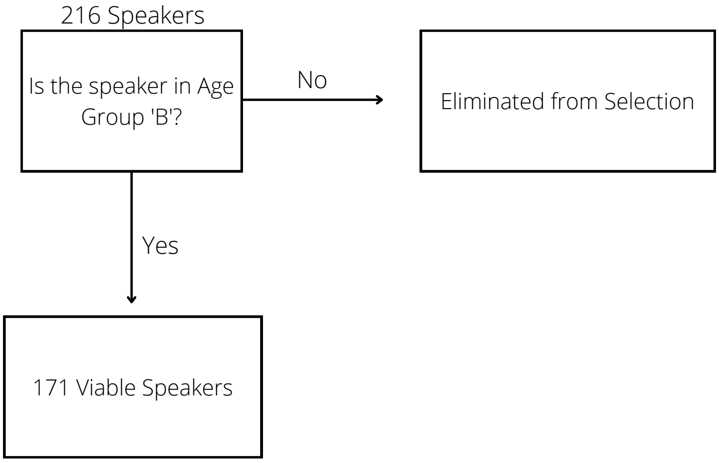


Fig. 6.2: Elimination of irrelevant age groups

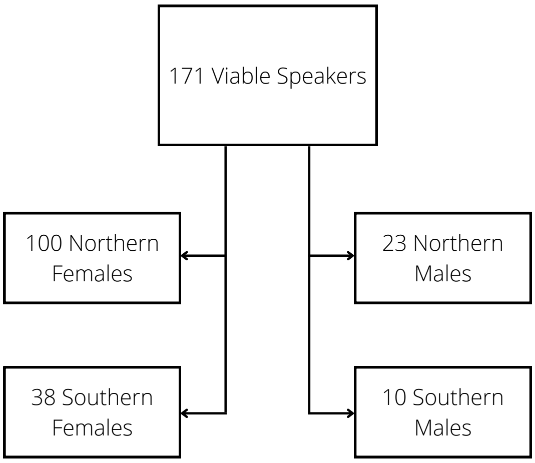


Fig. 6.3: Division into four sub-demographics

At this point in the file selection process, there were four clear divisions of speakers in the Excel sheet that correspond to the four demographic combinations in Fig. 6.3. The preliminary study investigated only /wei/ utterances from 5 speakers of each demographic. For this thesis, the intention is to drastically increase the number of speakers from the preliminary study (Crowley 2022). This is because the research question is inherently based on the differences between speakers of different demographics. By increasing the number of speakers studied, the possibility of taking a more representative sample increases accordingly. For this reason, the total number of speakers analyzed for this thesis will be 100, divided representatively among the four demographics, as to keep the sample comparable to what is available in the corpus. The numbers of each demographic combination studied in this thesis is shown in Table 6.3.

	NORTHERN	SOUTHERN
MALE	19	10
FEMALE	46	25

Table 6.3: Remaining number of speakers analyzed for this thesis from each demographic

To find which participants could be chosen from each of the demographic's totals, each speaker was checked for their own usages of the two sounds under investigation. To do this, the participant ID was checked against the total filtered utterances (Fig. 6.1). If there were at least 10 proper utterances, this speaker's sound files would be analyzed. If not, they were omitted and the next participant in the list was checked. In practice, there were very few total instances where a speaker had to be omitted due to a lack of appropriate utterances. The Southern male demographic required that some sound files be repeated due to almost every speaker in this demographic needing to be analyzed; there were not enough speakers in this demographic to be selective of which were chosen, even if one speaker had too few /wei/ or /wen/ utterances. All other demographics had excess speakers not initially chosen in the analysis, which could be pulled from if the first selected groupings did not contain enough appropriate utterances. Additionally, some individual sound files had to be replaced due to both test words being present. Fig. 6.4 was repeated for each of the 4 demographic combinations.

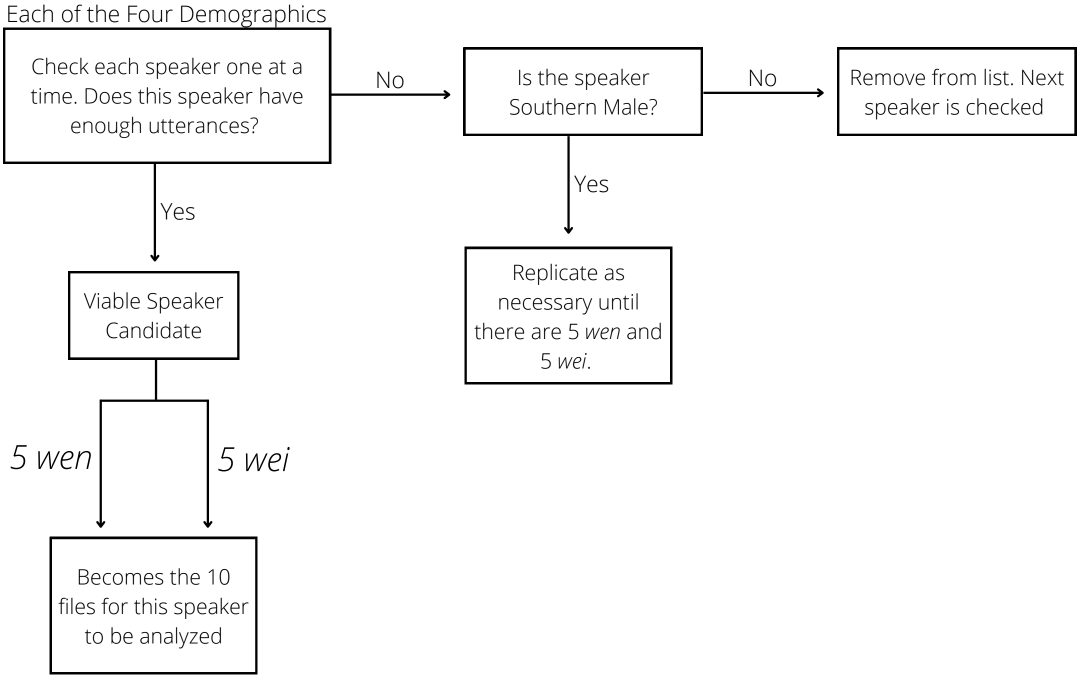

Each of the Four Demographics
Check each speaker one at a time. Does this speaker have enough utterances?
No
Is the speaker Southern Male?
No
Remove from list. Next speaker is checked
Yes
Yes
Viable Speaker Candidate
Replicate as necessary until there are 5 *wen* and 5 *wei*.
5 wen
5 wei
Becomes the 10 files for this speaker to be analyzed

Fig. 6.4: Final selection criteria for the total 1,000 sound files analyzed

Once all participants were found, their 1,000 sound files were placed into different sound folders in preparation for annotation and acoustic analysis via Praat (Boersma & Weenink 2022).

6.4 Annotation and Measurements

Only the initial /w/ of each sound will be analyzed. It is important that the annotation process remain as uniform as possible. Because of this necessity, the TextGrid boundaries had to be set in a systematic way that would not allow for inconsistent, researcher-oriented variation. The expected differences in F2 between [w] and [ʋ] were explained at the very end of § 4.3.1. and can be seen within the words /wei/ and /wen/ in Figs. 4.3 and 6.5. The dynamic range of the spectra were lowered to 30.0 dB if deciphering a boundary was difficult, but in practicality, this was almost never necessary.

The left boundaries were simple to select, as this sound only occurs as an onset or medial glide, depending on the rime structures. Therefore, this boundary was always chosen as the start of the visible F2 change. Sometimes, particularly when the articulation was a prominent [ʋ], there was no visible glide and F2 appears faint at the left side of the boundary, due to a large dip in intensity; this was exacerbated if the spectra's dynamic range was reduced as in the previous paragraph. This can be seen in Fig. 6.5. If this were the case, the left boundary was always chosen at the point where F2 became most visible. Because /w/ is only possible at the beginning of a syllable in *Pǔtōnghuà*, preceding words sometimes blended with the [w] onset, making precise boundary selection finicky, but there were no large errors that required a revamp of the selection procedures. The right-side boundary was selected with a similar thought process. What was deemed the best fit for this boundary was to choose a point after the visible glide where F2 stopped rising. No matter the strength of the articulation on the scale from [w] to [ʋ], there was almost always a visible point at which F2 plateaued; it was this point that became the right boundary. An example of both a strong [w] and [ʋ] can be seen in Figs. 6.5 below and 4.3 earlier.

The measurements that will be taken are also an important consideration. The following parameters are the measurable features that were deemed to be most relevant for studying this sound change. It is these features that are most indicative of and can provide the most reliable representation for this sound change. The reasons why each one was chosen will also be discussed below. The measurements are:

1. Lowest F2 Value
   - The difference between [w] and [ʋ] appear prominent with respect to F2. The literature, regardless of language, agrees that the patterns of F2 for these sounds are quite different. The F2 of the labiovelar approximant [w] begins lower and has a more prominent upward glide. The labiodental [ʋ] begins at a higher point.
2. Slope of F2
   - Because [ʋ] begins at a higher frequency, there is less vertical movement of F2. As a result, F2 changes significantly less over time than the [w] variant. In other words, the slope of [ʋ] is expected to be flatter. Duration values will not be analyzed, and instead are absorbed in the measure of F2 slope.
3. Harmonics-to-Noise Ratio
   - The articulation of these two sounds is similar but elicits a slightly different acoustic signal. Articulation of [w] is less constricted in the vocal tract, and as a result produces more laminar flow. Articulation of [ʋ] has an additional constriction at the front of the vocal tract via the teeth and lower lip, creating more turbulent airflow, and therefore more noise. This will lower the relative harmonicity of this sound. I expect [w] to have a higher HNR.

Examples of two spectrograms are shown below in Fig. 6.5. These spectrograms are from one of the chosen speakers from this paper. Red bars have been placed over these images to show roughly where both boundaries were placed in each sound file. It is between both red boundaries that all the aforementioned features were measured.

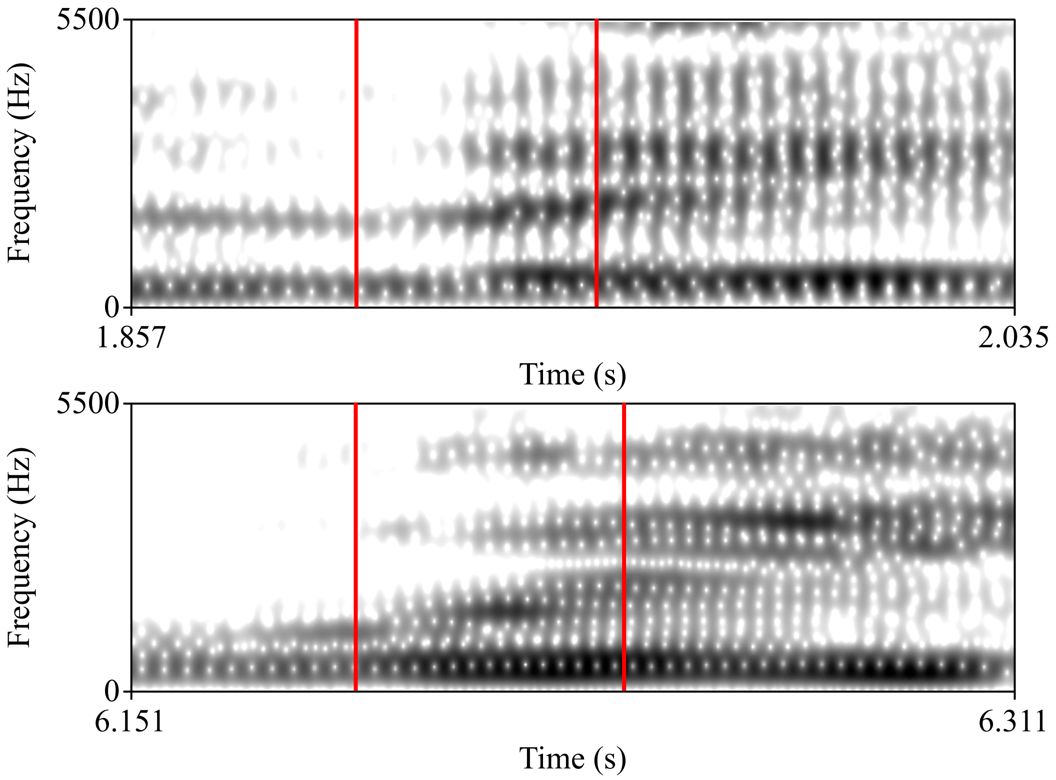


Fig. 6.5: Example spectrograms from the same speaker depicting average boundary locations. The top image is of a typical [ʋ] articulation. Bottom image is a typical [w] articulation.

To extract the desired measurements, a Praat script was run on all the sound files and TextGrid pairs. This script can be found in Appendix A. The first script was used to save the TextGrids together with sound files in a way that they would not become a praat.Collection file, as the data extraction script relied on having individual sound files and TextGrids in a directory of choice. In total, the extraction script was run four times, one for each speaker demographic. The data was copied into an Excel sheet, delimited to adjacent columns, then exported to a .csv file where it awaited statistical analysis in R.

7. Results

Statistical analysis was completed using R. The data were run through multiple different linear mixed effects models using orthogonal contrasts. These contrasts were swapped after the first model, which will be explained below. This chapter begins with a summary of critical results, followed by more detailed depictions of each model. Based on the methodology of this thesis, it is not possible to definitively tie these data to specific articulations, as acoustic measurements cannot make the inside of the speaker's mouth visible. However, these data are objective and clearly show trends that *do* suggest specific articulations with respect to each of the predictors (region, gender, and word). Lastly, this result section lays out which of the predictors had the strongest influence on the acoustic measurements.

In the subsequent sections of this chapter, all orthogonal contrasts were set up in a way that was consistent with the previous literature, regardless of my specific hypotheses. Each contrast will be clearly described just before the presentation of the statistical analysis for that contrast. The orthogonal contrasts were set up on either -0.5 or +0.5 when running the statistical analyses, depending on the expected trend of the data that was suggested by the literature, but are graphed on the contrasts -0.4 and +0.4, *only visually* to ensure the axes line up symmetrically. The subsequent sections will show each contrast as well as a graph depicting each result that was statistically significant.

A summary of all the results will now be mentioned along with a visual representation in Table 7.1. The gender predictor was most likely to produce a significant effect on the measurements as compared to region; there were fewer instances where a meaningful conclusion could be drawn from the region data. The results from the $F2_{min}$ data were largely inconclusive, as the effect of region turned out to go against my hypothesis, staying in-line with most of the established literature that speakers in the North have an $F2_{min}$ that is higher than those speakers in the South (Shen 1987; Wang 2011), implying that there is more usage of [ʋ] in the North. Additionally, the gender contrast for this measurement lay cleanly on an unsurprising anatomical difference between the genders: females had a higher F2 value.

However, the results from F2 slope and HNR were much more interesting. The results from F2 slope and HNR all suggest that male speakers may use this sound change more than females. The male speakers, on average, had a flatter F2 slope as well as a lower HNR. Region data for both F2 slope and HNR was not significant, so no conclusions can be reliably drawn.

Measurement	Contrast	Significant?	More [ʋ]	Hypothesis
$F2_{min}$	Gender	Yes	Female	Against
$F2_{min}$	Region	Yes	North	Against
$F2_{min}$	Word	Yes	/wei/	Against
F2 Slope	Gender	Yes	Male	In-Line
F2 Slope	Region	No	NA	NA
F2 Slope	Word	Yes	/wen/	In-Line
HNR	Gender	Yes	Male	In-Line
HNR	Region	No	NA	NA
HNR	Word	Yes	/wen/	In-Line

Table 7.1: Summary of (1) significant results from this thesis, (2) which member of the contrast group is suggested by that measurement to have more usage of [ʋ] and (3) alignment with hypotheses, as laid out in § 1.1.

7.1 Means and Standard Deviations

The following tables show the means and standard deviations of the 4 measurements, F2$_{min}$, F2 Slope, Duration, and HNR with respect to the 8 possible combinations of the 3 binary predictors, gender, region, and word. Some excess decimals were removed for the sake of clarity.

Table 7.2 shows the means from this experiment. Unsurprisingly, the values for F2$_{min}$ are all higher for female speakers than for the male speakers. This trend has been well documented before, but the difference in F2 frequency for one of these groups is interesting. Three of the four comparisons being held constant (north /wei/, south /wei/, and north /wei/) produced a range that clearly resembles the typical F2 difference across gender, between 150Hz and 220Hz. However, when north + /wen/ is held constant, this F2 difference between genders is significantly smaller (83Hz). This suggests (1) a possible increased usage of the labiodental variant [ʋ] in Northern male speakers compared to male speakers in other regions or (2) less [ʋ] usage in Northern females.

Additionally, this exact same trend of comparing two demographics, with the others being held constant, remains consistent when observing differences between region and word. Just as an example, Northern females produce a higher F2 than Southern females when the word is held constant and male speakers using the word /wei/ have a higher F2 when they are Northern than when they are Southern. It is only this first measurement, F2$_{min}$ that aligns itself with each demographic. The others align themselves with fewer demographics (Slope and HNR) or no demographic (duration). All these trends can be seen from Tables 7.2 and 7.3. Direct comparisons that were made in this paragraph have lines connecting the 2 values being compared in Table 7.2. A case will be made in the middle of § 8.2 for why measurements of duration are still present despite not being discussed at this moment.

Gender	**Region**	**Word**	**F2$_{min}$ (Hz)**	**Slope (Hz/ms)**	**Duration (ms)**	**HNR (dB)**
female	north	wei	1289	11.49	72.8	14.04
male	north	wei	1136	9.48	66.6	11.80
female	south	wei	1187	12.20	70.3	14.08
male	south	wei	1007	9.82	73.1	11.87
female	north	wen	1207	7.78	69.4	14.41
male	north	wen	1124	5.68	65.7	10.61
female	south	wen	1132	8.97	65.9	13.58
male	south	wen	918	6.54	78.1	10.70

Table 7.2: Means of each of the four measurements with respect to each demographic combination. Lines show the comparisons made in the previous paragraph.

Gender	**Region**	**Word**	**F2$_{min}$ (Hz)**	**Slope (Hz/ms)**	**Duration (ms)**	**HNR (dB)**
female	north	wei	295	4.86	15.7	4.24
male	north	wei	228	3.83	14.9	4.31
female	south	wei	247	4.78	18.4	3.83
male	south	wei	169	3.12	14.4	3.97
female	north	wen	241	4.72	14.5	4.87
male	north	wen	179	2.88	15.2	3.96
female	south	wen	206	4.28	15.5	3.99
male	south	wen	144	2.91	17.7	4.37

Table 7.3: Standard deviations of the four measurements with respect to each demographic combination

7.2 $F2_{min}$ Measurements

This parameter is a bit of a misnomer. Despite being described as the minimum F2 value, it is really taken as the F2 value at the left boundary location within the Praat annotations. This notation remains in place because it (1) efficiently conveys the effective use of "the lowest part of the signal" and (2) using the term $F2_{left}$ does not give vertical information as does using the subscript "min." In most cases, equating $F2_{min}$ with $F2_{left}$ resulted in no issues, as the boundary selection was done with respect to the visible glide, so the left boundary was almost always the lowest portion of F2 anyway. Using the F2 value at the left boundary was preferable because it (1) was less prone to automatic formant tracking errors that may occur at any point in the duration of the annotation, as Praat sometimes implants formant measurements where they do not exist and (2) choosing the left and right-side values of the annotation boundary to be the inputs for slope calculations created higher accuracy. Both reasons are justified via Fig. 7.1 below. Errors in formant tracking were still possible at the exact boundary locations, but in practice, due to the systematic nature of how the boundaries were selected, this rarely posed a problem.

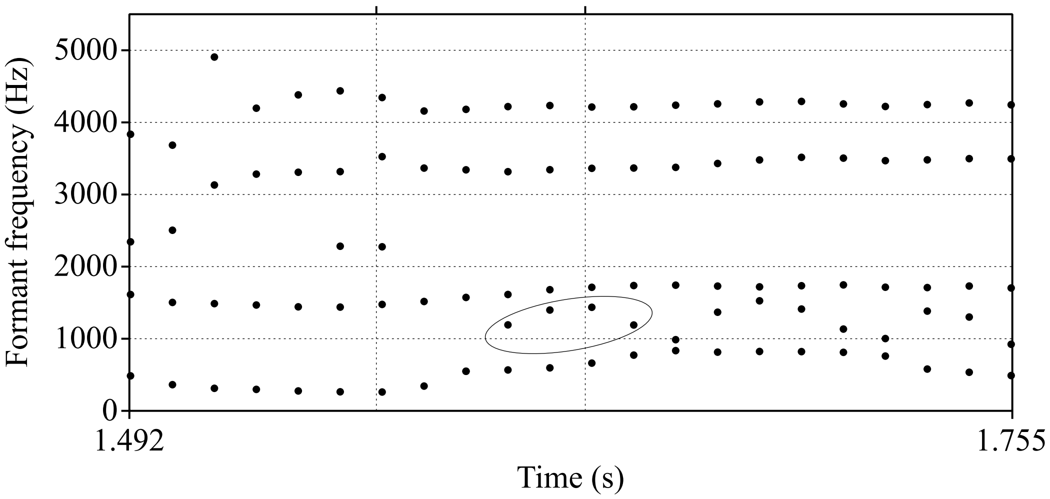


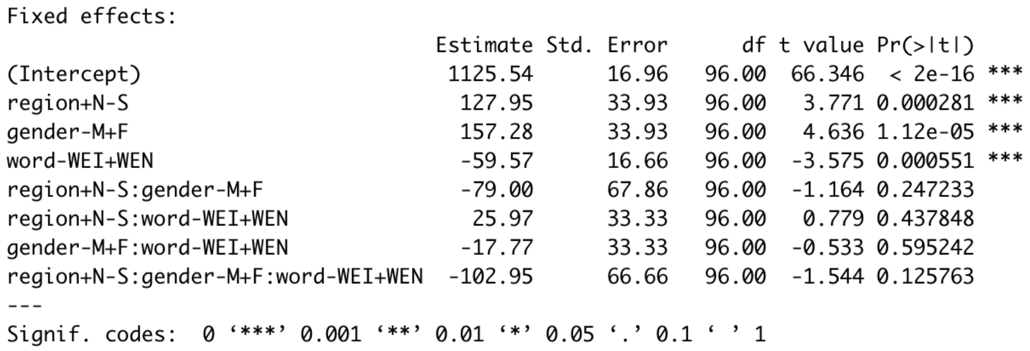
Fig. 7.1: A formant map of one speaker's utterance (SSB05020153) of /wen/ showing errors in automatic formant tracking within the boundaries (vertical lines). The dots for F1 appear consistent, but there is a string of auxiliary dots which would cause an error in tracking $F2_{min}$ as truly $F2_{min}$. Four of these misplaced dots have been circled.

7.2.1 Contrasts and Model for $F2_{min}$

The orthogonal contrasts for measuring $F2_{min}$ were set up as below. Why they are set up this way can be found in § 8.

- Gender: -M+F
- Region: -S+N
- Word: -WEI+WEN

A linear mixed-effects model was run on this dataset with these contrasts and outcome variable in place. The estimated linear intercept was 1126 Hz, which is the grand average of all F2 measurements, irrespective of group or contrast. The relevant portion of this LMER model can be seen in Fig. 7.2. Plots for significant effects are shown after the model summary.



```
Fixed effects:
                                       Estimate Std. Error     df t value Pr(>|t|)    
(Intercept)                             1125.54      16.96  96.00  66.346  < 2e-16 ***
region+N-S                               127.95      33.93  96.00   3.771 0.000281 ***
gender-M+F                               157.28      33.93  96.00   4.636 1.12e-05 ***
word-WEI+WEN                             -59.57      16.66  96.00  -3.575 0.000551 ***
region+N-S:gender-M+F                    -79.00      67.86  96.00  -1.164 0.247233    
region+N-S:word-WEI+WEN                   25.97      33.33  96.00   0.779 0.437848    
gender-M+F:word-WEI+WEN                  -17.77      33.33  96.00  -0.533 0.595242    
region+N-S:gender-M+F:word-WEI+WEN      -102.95      66.66  96.00  -1.544 0.125763    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 7.2: Cropped summary of the LMER F2$_{min}$ model.

The region contrast for the F2$_{min}$ model produced a statistically significant difference between speakers in the North and the South, and the average difference between North and South is 128 Hz, with a 95% confidence interval running from 62 Hz to 194 Hz, p = 0.00028. The positive coefficient means that based on these contrasts, the speakers in the North, on average, had an F2$_{min}$ value 128 Hz higher than the speakers in the South (Fig. 7.3). The data for the gender contrast are also significant, but this is not surprising. The estimated intercept for the gender contrast was 157Hz, with a 95% confidence interval for the running from 91 Hz to 223 Hz, p = 1.1E-05, suggesting that female speakers, on average, have an F2$_{min}$ value 157 Hz higher than males (Fig. 7.4); this will be discussed in § 8.1. Additionally, the model produced another statistically significant effect on the word contrast. There was an estimate of -60 Hz, with a 95% confidence interval running from -92 Hz to -27 Hz, p = 0.00055, suggesting that, on average, the word /wei/ produced an F2$_{min}$ value that was 60 Hz higher than the /wen/ words (Fig. 7.5). Full model summaries and confidence interval reports can be found in Appendix B.

The three significant contrast plots that were created for the F2$_{min}$ model are shown. Each side of the contrasts does have a different amount of data. The region contrast chart was aligned 650N:350S, the gender contrast chart was aligned 710F:290M, and the word contrast chart was aligned 500wei:500wen (not technically 1wei:1wen).

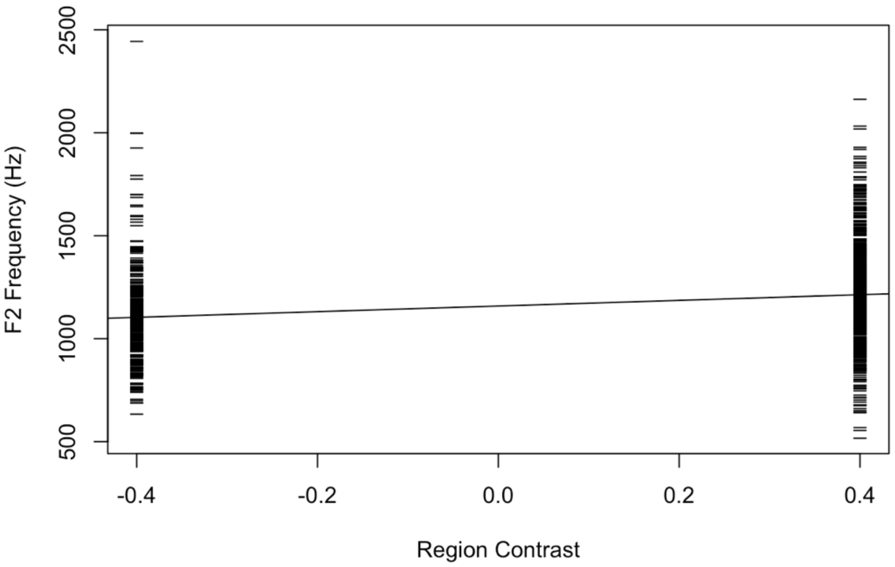


Fig. 7.3: South is coded as -0.4 and North is coded as +0.4 with an abline fitted across the two averages.

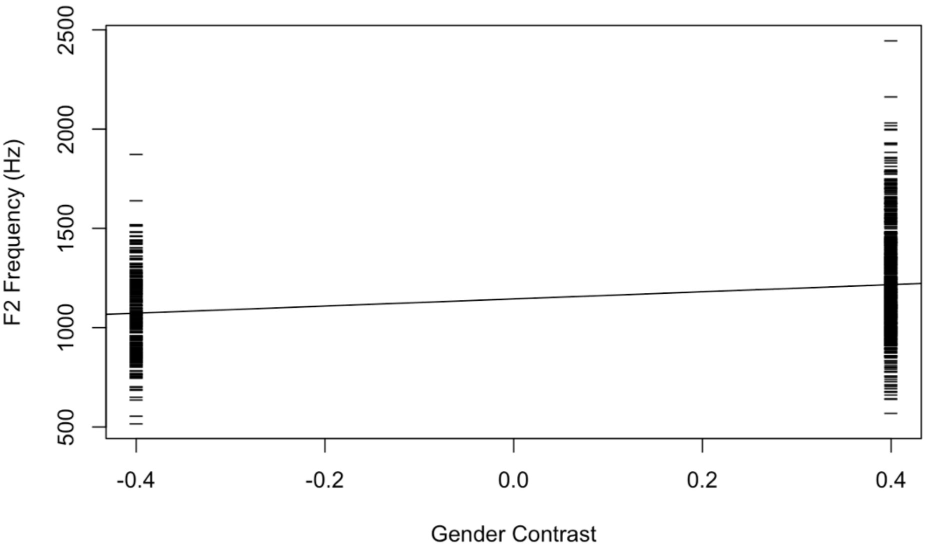


Fig. 7.4: Males are coded as -0.4 and female are coded as +0.4 with an abline fitted across the two averages.

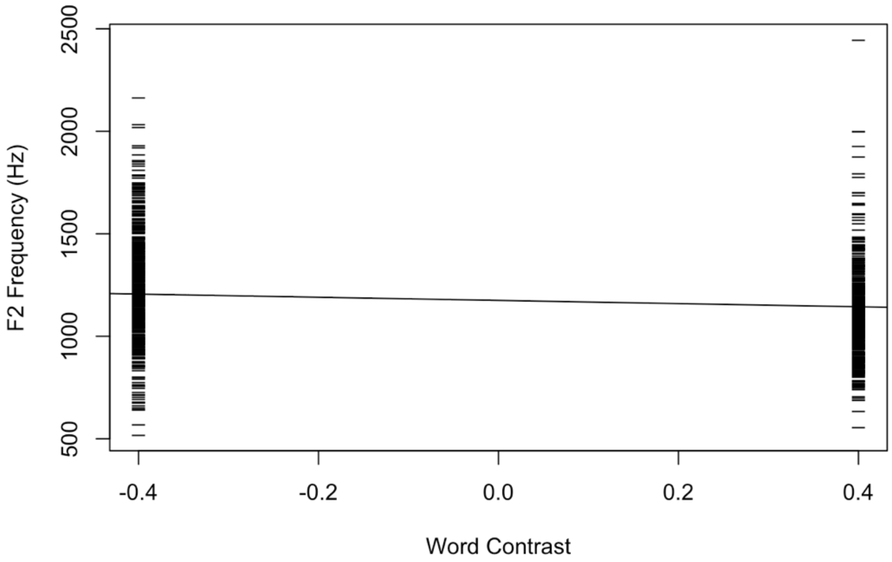


Fig. 7.5: /wei/ is coded as -0.4 and /wen/ is coded as +0.4 with an abline fitted across the two averages.

7.3 F2 Slope Measurements

Values for F2 slope were calculated directly within the Praat script, which can be viewed in its entirety in Appendix A. It was calculated using the F2 value at the left boundary subtracted from the F2 value at the right boundary, all standardized by the duration.

$$Slope = \frac{F2_{max} - F2_{min}}{Duration}$$

Equation Set 4: Calculation of F2 slope

despite the previous description on the discrepancy between $F2_{min}$ and $F2_{left}$.

Additionally, the term “F2 slope” is used to counteract possible confusion from the similarly named “slope” when discussing coefficient values in R models. “F2 slope” is used to mean the physical rate of change of F2 that is visible in a spectrogram, which has been described previously in §§ 3.1 and 3.2. If there is any mention of a coefficient slope, or a positive/negative direction within the coefficients of the model, there will be an obvious differentiation between this and the F2 slope.

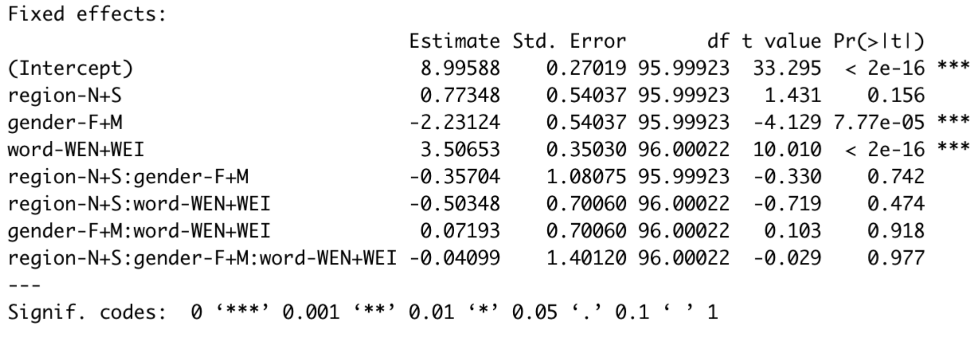
Finally, there is an important distinction for how the F2 slope contrasts are aligned. The subsequent contrasts for slope are expected to be *flatter* (smaller, `–`) or *steeper* (larger, `+`) to be used as an indicator of [ʊ] usage. At first, it may seem counter-intuitive to correlate a *lower* slope with *more* usage of [ʊ], but in terms of acoustic measurements and the specific hypotheses of this thesis, this is the most accurate way to describe it. This distinction is crucial before showing the contrasts for slope, as it is a compound measure, Hz/ms, not a measurement simply in Hz and it may be confusing to notate the F2 slope in this way.

7.3.1 Contrasts and Model for F2 Slope

The orthogonal contrasts for measuring slope were constructed in the opposite direction as for measuring $F2_{min}$. This is due to the inversely proportional relationship between gender and F2, which so happened to carry over directly to the other two contrasts. The new contrasts are as below. Why they are set up this way can be found in § 8.

- Gender: -F+M
- Region: -N+S
- Word: -WEN+WEI

A linear mixed-effects model was run on this dataset with these contrasts and outcome variable in place. The estimated linear intercept was 9.00 Hz/ms, which can be considered the grand average of all F2 slope measurements, irrespective of group or contrast. The relevant portion of this LMER model can be seen in Fig. 7.6.



```
Fixed effects:
                                    Estimate Std. Error       df t value Pr(>|t|)    
(Intercept)                          8.99588    0.27019 95.99923  33.295  < 2e-16 ***
region-N+S                           0.77348    0.54037 95.99923   1.431    0.156    
gender-F+M                          -2.23124    0.54037 95.99923  -4.129 7.77e-05 ***
word-WEN+WEI                         3.50653    0.35030 96.00022  10.010  < 2e-16 ***
region-N+S:gender-F+M               -0.35704    1.08075 95.99923  -0.330    0.742    
region-N+S:word-WEN+WEI             -0.50348    0.70060 96.00022  -0.719    0.474    
gender-F+M:word-WEN+WEI              0.07193    0.70060 96.00022   0.103    0.918    
region-N+S:gender-F+M:word-WEN+WEI -0.04099    1.40120 96.00022  -0.029    0.977    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 7.6: Cropped summary of the LMER Slope model.

The effect of region in this model is positive, producing an estimate of 0.77 Hz/ms, but did not produce statistical significance, with a 95% confidence interval that contains 0 and runs from -0.27Hz/ms to 1.82Hz/ms, p = 0.156. As a result, no true effect of region on slope can be concluded or extrapolated, even if the data leans slightly in one direction. However, the summary of this model provides a negative value for the estimate of gender; this result is opposite of what the literature would have predicted. The estimate for gender was -2.23Hz/ms, with a significance shown in a 95% confidence interval that runs from -3.28 Hz/ms to -1.18 Hz/ms, p = 7.77E-05, suggesting that, on average, males have an F2 slope value that is 2.23 Hz/ms flatter than females (Fig. 7.7). The fact that this estimate is negative is interesting and is discussed much further in the following discussion section. The effects of the word contrast are strikingly significant as well, and precisely in the direction predicted by the contrasts. The estimate generated by this summary is 3.51 Hz/ms, with a 95% confidence interval running from 2.82 Hz/ms to 4.19 Hz/ms, *p* < 2E-16, suggesting that the word /wei/ has, on average, a slope that is 3.51 Hz/ms steeper than the word /wen/ (Fig. 7.8). Full model summaries and confidence interval reports can be found in Appendix B. The following plots for the significant contrasts of the slope model are aligned exactly as described just before Fig. 7.3.

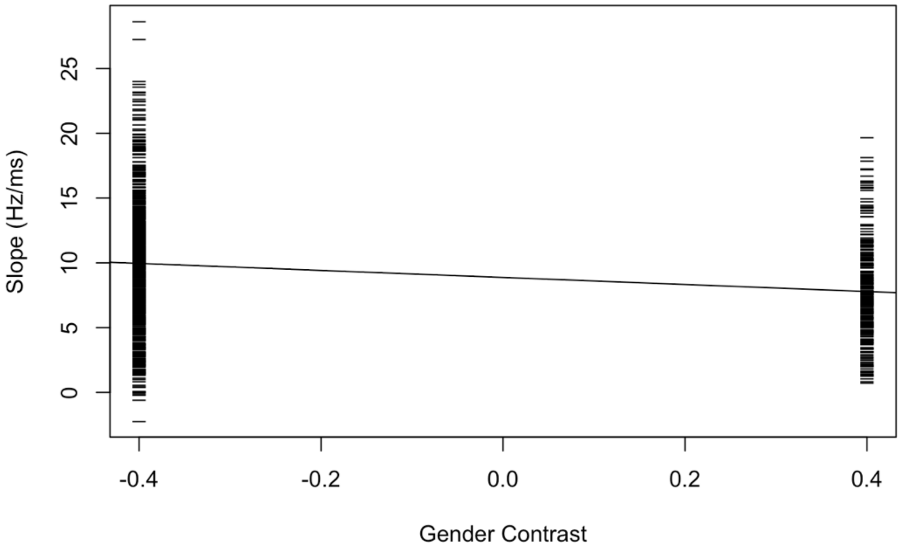


Fig. 7.7: Females are coded as -0.4 and males are coded as +0.4 with an abline fitted across the two averages.

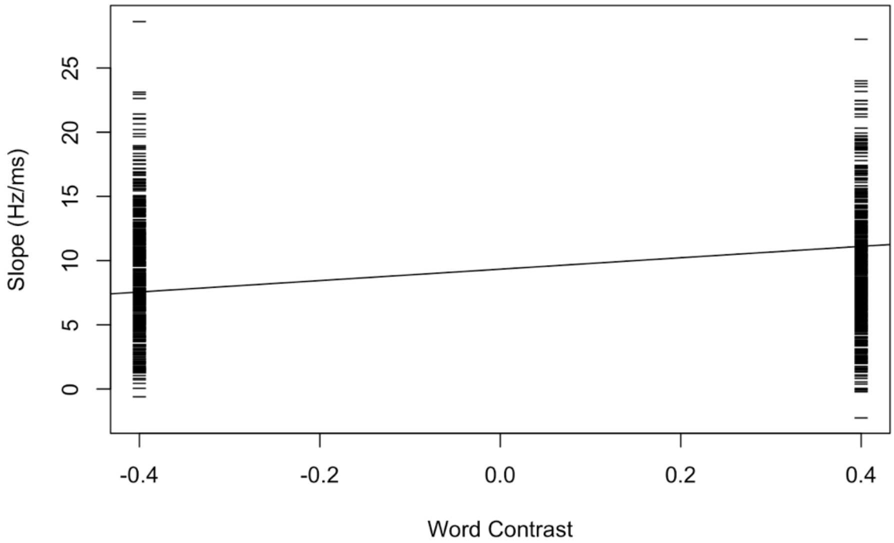


Fig. 7.8: /wen/ is coded as -0.4 and /wei/ is coded as +0.4 with an abline fitted across the two averages.

7.4 HNR Measurements

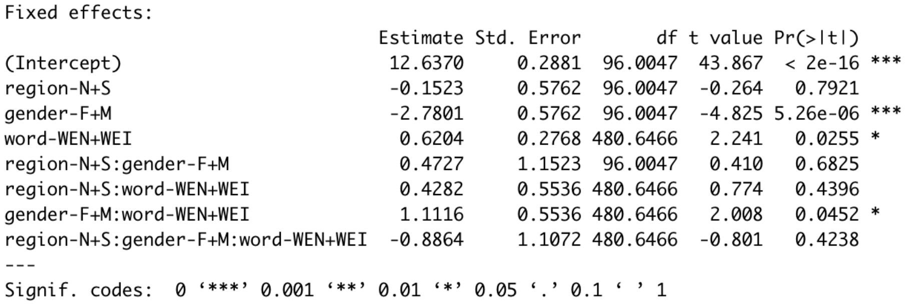
Measurements for HNR were recorded directly with the Praat script (refer to Appendix A). The sound segment was isolated and harmonicity was measured via the recommended cross-correlation method rather than the more technically sophisticated autocorrelation described in Boersma (1993). This thesis used a *harmonics-to-noise* ratio and not a *noise-to-harmonics* ratio (the Praat voice report lists both). A *lower* ratio indicates that there is more aperiodic noise while a *higher* ratio indicates that more of the energy in the signal lies in the periodic parts.

7.4.1 Contrasts and Model for HNR

The orthogonal contrasts for measuring HNR are identical to those used to measure slope. They are repeated here for convenience.

- Gender: -F+M
- Region: -N+S
- Word: -WEN+WEI

A linear mixed-effects model was run on this dataset with these contrasts and outcome variable in place. The estimated linear intercept was 12.64 dB, which can be considered the grand average of all HNR measurements, irrespective of group or contrast. The relevant portion of this LMER model can be seen in Fig. 7.9.



```
Fixed effects:
                                   Estimate Std. Error       df t value Pr(>|t|)    
(Intercept)                         12.6370     0.2881  96.0047  43.867  < 2e-16 ***
region-N+S                          -0.1523     0.5762  96.0047  -0.264   0.7921    
gender-F+M                          -2.7801     0.5762  96.0047  -4.825 5.26e-06 ***
word-WEN+WEI                         0.6204     0.2768 480.6466   2.241   0.0255 *  
region-N+S:gender-F+M                0.4727     1.1523  96.0047   0.410   0.6825    
region-N+S:word-WEN+WEI              0.4282     0.5536 480.6466   0.774   0.4396    
gender-F+M:word-WEN+WEI              1.1116     0.5536 480.6466   2.008   0.0452 *  
region-N+S:gender-F+M:word-WEN+WEI  -0.8864     1.1072 480.6466  -0.801   0.4238    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 7.9: Cropped summary of the LMER HNR model.

The effect of region on HNR was not significant, as the 95% confidence interval contains 0 and runs from -1.27 dB to 0.96 dB, *p* = 0.792. Therefore, it cannot be concluded whether or not there is truly an effect of region on HNR. However, gender did create a significant result, with this model producing an estimate of -2.78 dB. This estimate is negative, and therefore lies opposite of what was expected based on the gender contrast and literature. This effect's 95% confidence interval runs from -3.90 dB to -1.66 dB, *p* = 5.26E-06, suggesting that females will have, on average, an HNR value that is 2.78 dB larger than male speakers for this [w, ʋ] divide (Fig. 7.10). Additionally, the effect of word produced a statistically significant result in the direction that was expected based on the set contrasts. This model produced an estimate for the word contrast of 0.62 dB, with a 95% confidence interval running from 0.078 dB to 1.16 dB, *p* = 0.0255, suggesting that the /w/ in the word /wei/ produces an HNR that is, on average, 0.62 dB higher than for the word /wen/ and is more likely to indicate [ʋ] articulation (Fig. 7.11). Full model summaries and confidence interval reports can be found in Appendix B.

Finally, there is one more statistically significant result based on these data. This model produced a significant estimate for the *interaction* between gender and word of 1.11 dB, with a 95% confidence interval running from 0.028 dB to 2.20 dB, *p* = 0.0452, suggesting that, on average, the effect on HNR of the word /wei/ is 1.11 dB higher for males than it is for females as compared to the effects on HNR of the word /wen/ (Fig. 7.12). The following plots for the significant contrasts of the HNR model are aligned exactly as described just before Fig. 7.3.

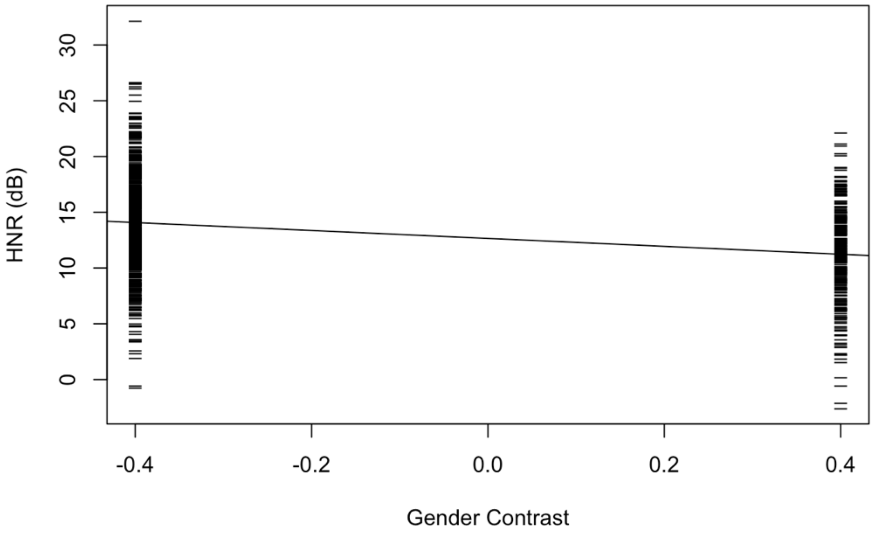


Fig. 7.10: Females are coded as -0.4 and males are coded as +0.4 with an abline fitted across the two averages.

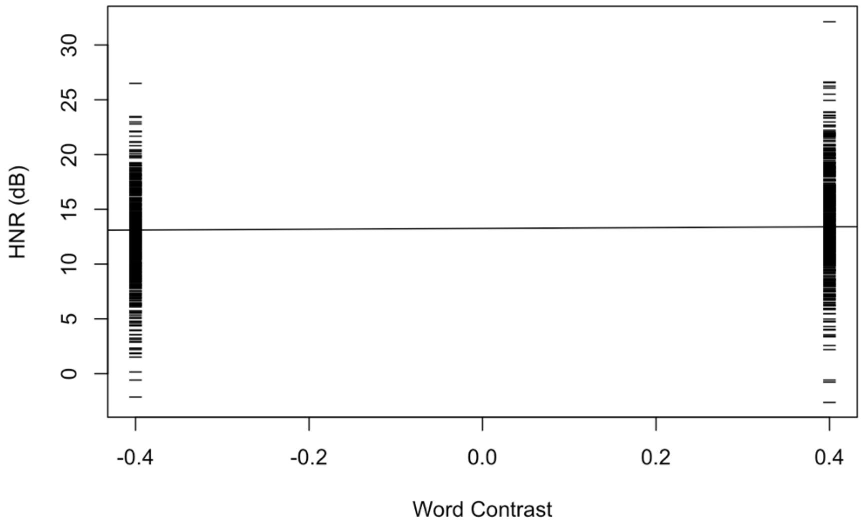


Fig. 7.11: /wen/ is coded as -0.4 and /wei/ is coded as +0.4 with an abline fitted across the two averages.

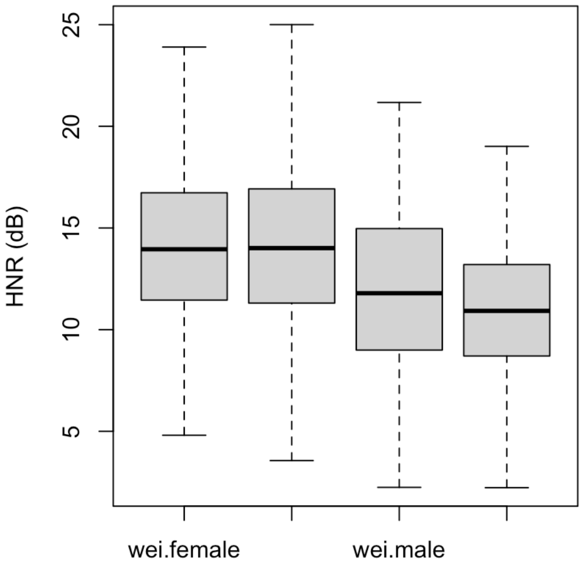


Fig. 7.12: Depiction of the interaction effect between the word and gender predictors for HNR (dB) measurement.

8. General Discussion

To briefly summarize the results of this thesis, some of the data could be extrapolated to suggest that males produce the labiodental sound change just as much as females, which falls opposite to what most of the previous literature would suggest (Hu 1991; Shen 1987; Chan 1996); this was true for the F2 slope and HNR measurements. This is a pivotal result, which has quite large implications on the distribution of sound change in China, especially since much of the established literature has been cited many times by authors from all over the world. However, where significance presents itself, some insignificance must be expected. Overall, the effects of region were much smaller than the effects of gender, no matter the measurement. In two of the three models, the effects of region cannot be properly substantiated due to statistical insignificance. Overall, this thesis brings to light some interesting trends that will now be elaborated upon in the following discussion.

This chapter is separated into four subsections, each dealing with one of the measurements taken during this experiment. First, a recap and elaboration of the chosen contrasts will be provided as well as deeper ties to the specific literature that were alluded to in making these contrasts. It has been said several times that the contrasts were all constructed in a way that was indicative of what the literature clearly states, or what can be extrapolated from the literature, even if it lied opposite to the specific hypotheses of this thesis. The contrasts will be repeated here now.

$F2_{min}$ Contrasts:

- Gender: -M+F
- Region: -S+N
- Word: -WEI+WEN

The *gender* contrast is in line with the well-established literature that males produce, on average, an F2 that is significantly lower than that of females (Reetz & Jongman 2009). The *region* contrast is in line with the literature that suggests the sound change from [w] to [ʋ] is more common in the North (Hu 1991; Chan 1996a), and since [ʋ] produces a higher F2 (Wiener & Shih 2011), the data for North speakers would be higher. The contrast for *word* is in line with literature stating that /wen/ produces more [ʋ] than [w] compared to /wei/ (Wiener & Shih 2011; Chao 1927) and will thus have a higher F2 frequency.

The contrasts for F2 slope and HNR are flipped from what they were for the $F2_{min}$ model and are repeated here now.

F2 slope and HNR contrasts:

- Gender: -F+M
- Region: -N+S
- Word: -WEN+WEI

The *gender* contrast here is based on the literature that females use [ʋ] more frequently (Hu 1987; Shen 1987), therefore causing a flatter slope. The *region* contrast is similarly based on the same literature as the $F2_{min}$ – region contrast, positing a flatter F2 slope due to increased usage of [ʋ] in the North. Finally, the *word* contrast is based on the same papers as before (Wiener & Shih 2011) that concluded /wen/ produces more [ʋ] than [w] compared to /wei/.

8.1 F2*min* Discussion

The results for the $F2_{min}$ model generally fell out of line with all my hypotheses, but there are still some interesting trends to point the discussion toward. The coefficient slopes for the estimates of region and gender are unsurprising, but what might be interesting, however, is the distance between the region and gender estimates. From the summary provided in § 7.2.1, it suggests that the difference in effect of both gender and region on $F2_{min}$ is only 29 Hz, suggesting that region has a similarly strong effect on $F2_{min}$ that gender does; gender's effect was still stronger with a slightly smaller p-value. Additionally, both coefficient slopes are positive, which means northerners and female speakers produce an $F2_{min}$ that is significantly higher than southerners and male speakers. However, this brings up an interesting point. It may be expected that since the effect of both region and gender are statistically significant (and to such a large extent), the effect of the interaction between gender and region must be as well. This makes sense superficially, as the most common demographic to utilize this sound change were female speakers who live in the North, but comparing these significance values by themselves is meaningless and it is not necessarily true that an interaction between predictors becomes significant based simply on the significance of the individual predictors (Gelman & Stern 2012). This comes up again at the end of § 8.3.

Despite not being in-line with the hypotheses of § 1.1, only the word contrast was unexpected. It is still impossible to determine conclusively just from $F2_{min}$ that northerners and female speakers are, in fact, utilizing the [ʋ] variant more than [w]. The only thing these data can show is the trend in that direction, from which an inference can be made, which will have other data analyses built onto it in the subsequent sections. Finally, a significant effect of gender on $F2_{min}$ could simply be the result of physiological differences between males and females.

The effect of word in this $F2_{min}$ model lies opposite to both the hypotheses and previous literature. This thesis expected the word /wen/ to have a higher F2 based on literature suggesting /wen/ was the onset + rime combination that resulted in the most labiodental sound change (Wiener & Shih 2011; Shen 1987). These studies indicated that the schwa-like vowel in /wen/ produced the most [ʋ], but that /wei/ was right behind. This summary shows a negative estimate for my word contrast, which implies that within the scope of this thesis, the word /wei/ had a higher $F2_{min}$ value, meaning that it might in fact be the set of words that produces the labiodental variant [ʋ] more than the /wen/ words. Future research will be needed to confirm or deny this.

Finally, according to these data, none of the interactions effects between groups were significant with respect to the $F2_{min}$ model; it is not possible to draw any conclusion on how the interactions between gender, region, and word affect $F2_{min}$ values.

8.2 F2 Slope Discussion

As mentioned at the end of § 7.3, the contrasts for F2 slope do not measure a single unit like $F2_{min}$. As a result, when viewing the coefficients from the model summaries, it is important to take this into consideration, as it may seem counter-intuitive to depict the one expected to have a *flatter* slope as the negative contrast.

Half of the results from slope are quite surprising, and half of them were as expected. First, the region contrast did not produce statistical significance, but the p-value was not wildly off from significance. The positive coefficient for the region estimate (if significant) would be in-line with the previous literature, but a subsequent study will need to be done to confirm that. The effect of gender in this analysis brings up something interesting, and this is where one of the largest discrepancies between the previous literature and the current hypotheses lie. As mentioned several times, the contrasts were created in a way that was in-line with the established literature that female speakers produce more [ʋ] than men (Hu 1987; Hu 1991). However, part of my hypotheses was that males will utilize [ʋ] either as much as females, or at least significantly more than previously suggested. The only piece of literature that posits the presence of [ʋ] in males to be significant was Zhou (2003). The negative coefficient in Fig. 7.6 is a strong indicator that males may use [ʋ] more than females. Females being left-coded in the contrasts creates this negative coefficient slope. Additionally, it is by quite a large margin as well. The statistical significance for this is ironclad, with such a low p-value (7.77E-5), suggesting that this is not simply a mistake in some measurements. The general trends of how slope aligns itself across the genders is corroborated by Table 7.2 earlier; the averages of F2 slope are visible in this table, and in each demographic combination, the males produced a smaller (flatter) slope.

Additionally, the F2 slope values obtained in this thesis were doubly standardized. This standardization removes the possibility that these F2 slope values are skewed for the gender contrast simply due to physiological differences between males and females. The measurements of slopes rely on multiple F2 measurements at different times and is then standardized over the duration of the utterance. The following equation shows how the differences between the F2 values can be the same, even if they lie at different frequency ranges, and how the only defining feature that can affect the values of the F2 slope ends up being duration.

$$\frac{2400Hz - 1600Hz}{70ms} = \frac{1750Hz - 950Hz}{70ms} = 11.43Hz/ms$$

Equation Set 6: Example calculations of F2 slope. Despite falling at different F2 ranges, they can produce identical slopes.

These values produce a difference of 800Hz prior to standardizing with duration. These variations in duration might have some effects on slope in isolated cases, but as a whole across the entire dataset used in this thesis, it is unlikely that there was an effect. As can be seen from Tables 7.2 and 7.3 column 6, some female demographics are higher and some male demographics are higher. Further, the standard deviations for duration are all quite similar.

Finally, the word contrast for F2 slope strongly aligned itself with the previous literature and the hypothesis. It was expected that the word /wen/ would have a significantly flatter slope (Wiener & Shih 2011), and the coefficient estimate corroborates this. A vast difference in slope is one of the defining features of the sound change from [w] to [ʋ]. In other words, the data strongly support the statement that /wei/ has a drastically steeper slope, especially with such a strong p-value. None of the interaction effects for the slope model were significant. None of them seem reasonably close to being significant, either. Therefore, it is impossible to tell based on these data the true interaction effects between any of the three predictors for *Pǔtōnghuà*.

8.3 HNR Discussion

The harmonics-to-noise ratio measurement is a bit nebulous of a parameter to measure. As described in § 3.3, it is the ratio of how much energy is stored in the periodic parts of a sound sample as compared to the aperiodic parts. The assumptions that were made for this thesis were that the articulation of the [ʋ] variant creates more turbulent airflow than the more standard [w] variant. Production of [w] only uses a single constriction, along the center of the vocal tract length, as can be seen in half of Fig. 4.1, repeated here as Fig. 8.1.

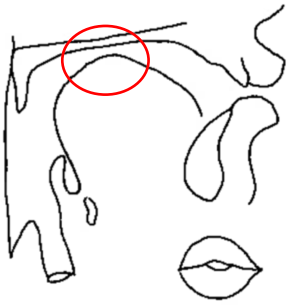


Fig. 8.1: Cross-sectional area of articulation of the sound [w] (Stevens 1998: 516)

However, the production of the [ʋ] sound involves an additional constriction of airflow at the front end of the vocal tract; this interaction between the lower lip and the upper teeth, even if it's very small in some speakers, will inevitably alter the output acoustics of this sound. With a [ʋ] articulation, the amount of noise present in the sound sample will be higher. Boersma (1993) and the Praat user manual define HNR mathematically as in the following equation:

$$HNR = 10 \times \log_{10}\left(\frac{Periodic\ Energy}{1 - Periodic\ Energy}\right)$$

Equation Set 7: Calculation and definition of HNR

If 95% of the energy in a signal is found within the periodic portion of the sound, the resultant HNR would be 12.79 dB. However, one important note that was described on the harmonicity page of the Praat user manual was the tendency for the vowel [u] to generally trend toward a lower HNR, mostly due to the higher harmonic frequencies that are present in more fronted or lowered vowels like [i] or [a]. Figure 8.2 show spectral envelopes for two vowels under question. Notice how the [a] vowel has much more energy throughout the frequency range, but the [u] vowel's energy is mostly limited to the lower range.

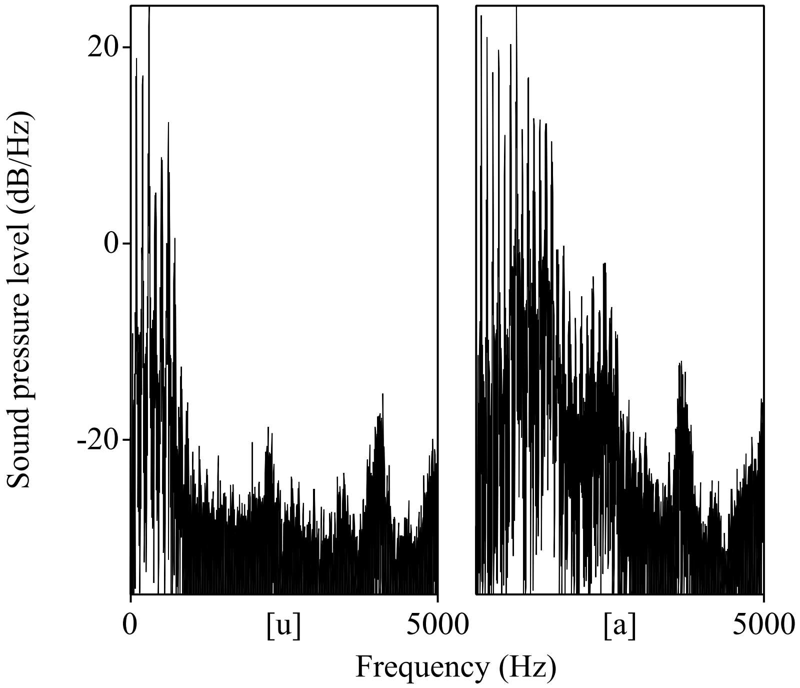


Fig. 8.2: Spectral envelope for production of a prototypical [u] vs. a prototypical [a] articulation

It appears that females were more likely to produce sounds that had a higher percentage of periodicity, possibly suggesting that they use the [w] variant more than males do. This result was opposite of what the previous literature would have predicted. The negative coefficient estimate of -2.78 dB places females as having a higher HNR, which is largely unexpected based on typical usage of [w] vs. [ʋ] across genders found in the literature (Hu 1991; Hu 1987; Shen 1987). However, this is yet another piece of evidence from this thesis that does reinforce the previous hypotheses that this sound change is widely prominent in male speakers as well. However, it must be noted that this may possibly be due to a physiological difference between males and females, but I am inclined to suggest that this is not the case. Since males produce an F0 that is generally lower in frequency, their harmonics will be closer together; this allows a higher number of harmonics to be present in a specific time frame as compared to females. As a result, I would expect that males broadly would have more energy present in the periodic portions of their speech signals, and thus a higher HNR. However, the data here strongly sugest that, even if that is accurate, males are *still* producing [] enough to overcompensate for this. I understand that all data should be taken with a grain of salt, but this is a very interesting trend. Lastly, the default setting for HNR extraction in Praat only identifies the voiced portions of the sound signal. It is possible that the measurements were simply inaccurate or had some sort of flaw in them, causing these data to be unreliable, but since the entire isolated sound segments were voiced, this may not be the case either. Further research may be needed to elaborate on this point.

Next, the trend for the word contrast on this model is in line with previous literature as well as my hypotheses. The positive estimate coefficient for this contrast supports the idea that the word /wen/ utilizes the labiodental variant [ʋ] more than the word /wei/, but the extent to which is not something that can be determined based on these statistical analyses. Lastly, the interaction effect of both word and gender on HNR is much less straightforward to interpret than the other effects that involved only a single predictor. It is important to make a note that these values are not just meant to explain the difference between two parts of a group; it is meant to show the variation between two different groups with respect to another group. The only significant interaction effect in this thesis (gender & word interaction) can be used to show a comparison between females and males when the two different words (/wei/ and /wen/) are spoken.

8.4 Additional Commentary and Error

The detriment to working with corpus data is that there is no real control by the researcher in what types of data can be used. Unfortunately, there was no way to control for specific measurements and words in a study of this nature; I had to rely only on what was available within the corpus, even if I could be selective for certain factors within the corpus itself. A follow-up study to this that does not have a dependency on a corpus ought to create a highly specific script for participants to read in order to precisely elicit the relevant information. Things such as preceding sound should also be controlled for, as that could influence the measurements recorded in this thesis. All the data for this thesis was properly recorded and analyzed, but there will always be a certain degree of inaccuracy associated with a study of this type with respect to how extrapolatable the results are.

Next, acoustics alone might not provide the most accurate information on physical usage of one sound over the other, even though it is the most objective series of measurements that can be taken on the subject. In determining true production information, things such as articulation studies and especially auditory perception experiments need to be run. A person may have an F2 pattern that closely resembles a prototypical [ʋ] acoustically, but it still may just *sound* like a true [w]. Fig. 8.3 below is of one of the speakers encountered during annotation in which the acoustics clearly resembled a [ʋ] pattern acoustically, but auditorily sounded much closer to [w], both to a native Chinese speaking colleague and myself.

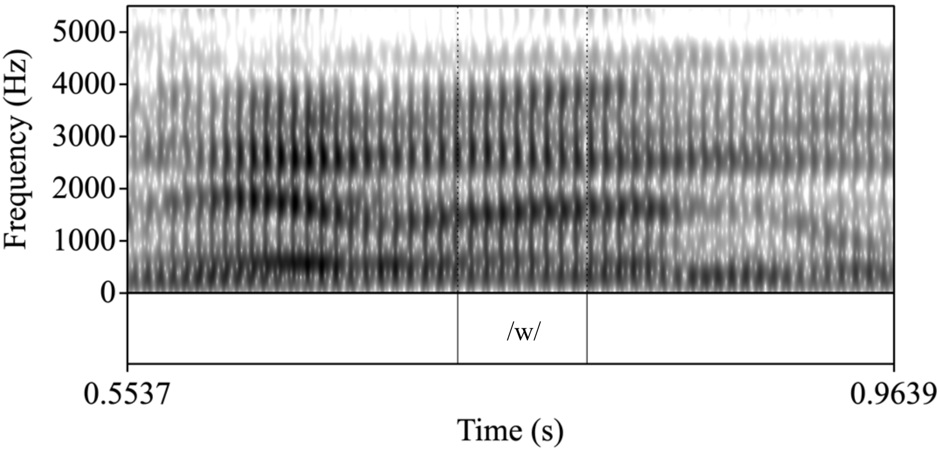


Fig. 8.3: Spectrogram and TextGrid of speaker SSB05440099. This speaker sounded like they used [w] but does not present acoustically as being a true [w]. However, examples of this were quite rare for this thesis.

These are the types of juxtapositions between sub-phonetic disciplines that are very important to take into consideration. Auditory information was not included in this thesis because the focus was primarily on acoustic and articulatory information, which should provide just as accurate a depiction of what speakers do. I would have liked to include auditory judgments in my analysis, but it was not feasible given the time frame of this thesis as well as how subjective auditory judgments can be.

Next, it is important to note minor changes to methodology in any way in the name of transparency. Most speaker and file selection, as described in chapter 6 was indeed random, except for a few cases where it may have been detrimental to rely on true randomness. In all instances when a speaker who was in line to be analyzed did not have enough /wei/ or /wen/ utterances, they were replaced with another speaker who *did* have enough of those words. This practice only had an effect on the Southern male speakers because replacement was not possible for this demographic, as there were so few speakers that all 10 of them were obligatorily included in the analysis. The speakers that ended up being replaced did not affect the analysis, but I believe it is important to mention that there was some degree of file reduplication in the Southern male demographic. Additionally, some sound files contained both the words /wei/ and /wen/, so in order to avoid overlapping, a sound file was either (1) skipped if it contained both test words or (2) kept, but only one of the words was analyzed, with an additional file replacing the second word. In total, 40 out of 1,000 total sound files were replaced with a backup in some way.

Finally, a future study with a similar research question ought to measure spectral intensity as well as take into consideration different age groups. Throughout the annotation process, there appeared to be a quite prevalent dip in intensity when some speakers used the labiodental variant [ʋ] instead of the standard labiovelar variant [w]. This dip in intensity can be seen in Figs. 4.3 and 6.5, as the lighter portions just to the left of the visible F2 glide, which signal the onset of the [ʋ] articulation. This may be an important characteristic in determining definitive usage of [ʋ] vs. [w], especially if the relationship between this sound change and intensity is more regular or patterned than perhaps some other metrics. Additionally, much of the seminal work on this sound change, specifically with respect to the predictors of region and age, took place over 30 years ago. The most recent study of this sound change (Wiener & Shih 2011) did not take into consideration the variations in age over time. Some important questions were introduced in § 5.3, and further research should be performed to answer these questions about the effects of age and time frames on this sound change.

9. Conclusion

This thesis investigated the prevalence of the sound change of the labiodental approximant [ʋ] in non-expectant demographics in *Pǔtōnghuà*. Data presented here suggests that male speakers may be using this sound change just as much as, if not more than females. However, all interpretations of data used in this thesis were extrapolatory; I aimed for this thesis to be accurate not only in terms of the phonetics, but also in terms of statistical analysis. Because of this, direct comparisons between acoustic measures and conclusions of articulation would not technically be correct or elicit accurate statistics. The results were presented as the differences in effects of predictors (region, gender, and word) on the three measurements ($F2_{min}$, F2 slope, and HNR). These differences tell us which of the predictors will have the strongest relative effects on our measurements. It is this comparison of differences that is extrapolated toward "usage of [ʋ]" based on the previous literature cited throughout this thesis. It is for this reason that statements suggesting more usage of [ʋ] in one group versus another is possible, whether that be the male-female or North-South divide.

The measurements of $F2_{min}$ were somewhat inconclusive, as the trends that arose from that LMER model either went against the hypotheses stated in § 1.1 (as was the case for the $F2_{min}$ – region contrast) or provided unsurprising information (as was the case for the $F2_{min}$ – gender contrast). Measurements for F2 slope and HNR all strongly suggest that the usage of [ʋ] in males is comparable to females.

The results of this thesis are novel; a study of this kind has, at least to my knowledge, never been conducted. The fact that this thesis very strongly shows evidence contrary to the large pool of established literature on the topic is worth recognition. More recent studies involving the labiodental approximant in Chinese (Wiener & Shih 2011) are focused on a slightly different research question. The conglomeration of all available research on the topic will only increase what future researchers can specify with their respective works. In §§ 5.3 & 8.4, I suggested possible places where future research could be developed from this study. Future researchers ought to further uncover the inner machinations of the production of this sound change within China and abroad, whether that be focusing on age variations over time, language contact and spread, spectral intensity measurements, or many others. Overall, this thesis brings to light some interesting trends that allow this work to stand on its own in the broader range of Chinese phonetics and sound change literature that I hope will be further developed in the future.

References

Anderson, S. (1976). On the description of multiply-articulated consonants. *Journal of Phonetics* 4, 17-27.

Ballew, R. (2019). Producing a Sound with Source-Filter Theory. Image from video: https://www.youtube.com/watch?v=Mj53hfwC9wQ

Basbøll, H. (2005). *The Phonology of Danish*. Oxford University Press.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the hermonics-to-noise ratio of a sampled sound. *Intitute of Phonetic Sciences*, University of Amsterdam. *Proceedings* 17, 97-110.

Boersma, P. & Weenink, D. (2022). Praat: doing phonetics by computer [Computer program]. Version 6.2.14, revrieved 24 May 2022 from http://www.praat.org/

Booij, G. E. (1999). *The Phonology of Dutch*. Oxford University Press.

Chan, K. M. (1996a). Sound symbolism and the Chinese language. In Cheng, T., Li, Y., Zhang, H., (Eds.) *Proceedings of the 7th North American Conference on Chinese Linguistics (NACCL) and the 4th International Conference on Chinese Linguistics (ICCL)* 2, 17–34. GSIL Publications.

Chan, K. M. & Lin, Y. (2019). Chinese Language and Gender Research. In Huang, C., Jing-Schmidt, Z., & Meisterernst, B. (Eds.) *The Routeledge Handbook of Chinese Applied Linguistics*, 165-181. Routeledge.

Chao, Y. (赵元任) (1927). *A New Vocabulary of Rimes* (*国音新诗韵*).

Chao, Y., Luo, C., & Li, F. (赵元任，罗常培，& 李方桂). (1937). Translated notes on Gao Benhan's *Research on Chinese Phonology* (中国音韵学研究的中译本). Referenced from inside Hu, M. 1987.

Chen, F. (陈凡凡). (2010). [v] Usage Frequency in Pǔtōnghuà and Tendency of Accepting [v] Among Chinese Speakers (汉民族共同语中【v】的使用倾向和接受度调查). *Shantou University Bimonthly Journal of Humanities and Social Sciences* (*汕头大学学报，人文社会科学版*) 26(1), 67-71. Accessed via Baidu (百度).

Chen, P. (1999). *Modern Chinese: History and Sociolinguistics*. Cambridge University Press.

Chiba, T. & Kajiyama, M. (1941). *The Vowel: Its Nature and Structure*. Phonetic Society of Japan.

Colombo, I. (2015). On the phonemic status of labial approximants in Dutch. rMA Thesis, University of Amsterdam.

Crowley, J. (2022). A Corpus Study on the Sound Change of the Voiced Labiodental Approximant [ʋ] in Mandarin Chinese. Unpublished manuscript of preliminary study. University of Amsterdam, Graduate School of Humanities.

Duanmu, S. (2000). *Phonology of Standard Chinese*. Oxford University Press.

Fant, G. (1960). *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*. (Second Printing 1970). Mouton. DOI: //.1515/9783110873429

Fernandes, J. et al. (2018). Harmonic to Noise Ratio Measurement – Selection of Window and Length. *Procedia Computer Science* 138, 280-285.

Garrett, A. & Johnson, K. (2011). Phonetic bias in sound change. In Yu, A. C. L. (Ed., 2013, pp. 51-97) *Origins of Sound Change*. Oxford University Press.

Graddol, D. & Swann, J. (1989). *Gender Voices*. Basil Blackwell.

Hu, M. (胡明扬)(1987). Investigation of the feminine accent in the Beijing vernacular (北京话‘女国音’调查). *China National Knowledge Infrastructure* (中国知网), 26-31. China Academic Journal Electronic Publishing House. DOI: 10.16412/j.cnki.1001-8476.1988.01.014

Hu, M. (胡明扬). (1991). Feminine Accent in the Beijing Vernacular: A Sociolinguistic Investigation. *Journal of the Chinese Language Teachers Association* 26(1), 49-54.

Hyman, L. (1976). Phonologization. In Jullian, A. D. (Ed.) *Linguistic studies presented to Joseph H. Greenberg.* Saratoga: Anma Libri, 407-418.

Hyman, L. (2013). Enlarging the Scope of Phonologization. In Yu, A. C. L. (Ed., 2013, pp. 3-28) *Origins of Sound Change*. Oxford University Press.

Institute of Sound and Vibration Research. Animations of Acoustic Waves. University of Southampton. http://resource.isvr.soton.ac.uk/spcg/tutorial/tutorial/Tutorial_files/Web-basics-nature.htm

Johnson, K. (2012). *Acoustic and Auditory Phonetics* (3rd Ed.). Wiley-Blackwell.

Kiparsky, P. (2015). Phonologization. In Honeybone, P. & Salmons, J. (Eds.) *The Oxford Handbook of Historical Phonology*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199232819.013.017

Ladefoged, P. (1971). *Preliminaries to Linguistic Phonetics*. University of Chicago Press.

Lehiste, I. & Peterson, G. (1961). Transitions, Glides, and Diphthongs. *The Journal of the Acoustical Society of America* 33 (3), 268-277.

Lin, T. (林焘). (1982). On [v] in Pǔtōnghuà (论普通话的【v】). *China National Knowledge Infrastructure* (中国知网), 1-5. China Academic Journal Electronic Publishing House.

Liu, J. et al. (2020). Study on the Spatial Differentiation of the Populations on Both Sides of the “Qinling-Huaihe Line” in China. *Sustainability* 12, 4545. DOI: 10.3390/su12114545

Maddieson, I. & Disner, S. F. (1984). *Patterns of Sounds*. Cambridge University Press.

Norman, J. (1988). *Chinese*. Cambridge University Press.

Ohala, J. (1983). Chapter 9: The Origins in Sound Patterns in Vocal Tract Constraints. *Production of Speech*, 189-216. Springer.

Ohala, J. & Lorentz, J. (1977). The Story of [w]: An Exercise in the Phonetic Explanation for Sound Patterns. *Proceedings of the 3rd Annual Meeting of the Berkeley Linguistics Society*, 577-599.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ramsey, S. R. (1987). *The Languages of China*. Princeton University Press.

Recasens Vives, D. (2014). *Coarticulation and Sound Change in Romance*. John Benjamins Publishing Company.

Reetz, H. & Jongman, A. (2009). *Phonetics: Transcription, Production, Acoustics, and Perception*. Wiley-Blackwell.

Shen, J. (沈炯). (1987). Phonetic differences of zero initial before finals beginning with u in the Beijing dialect (北京话合口呼零声母的语音分歧). *Chinese Language* (*中国语文*) 5, 352-362.

Shi, Y. et al. (2020). AIShell-3: A multi-speaker Mandarin TTS Corpus and the Baselines. https://arxiv.org/abs/2010.11567

Stevens, K. (1998). *Acoustic Phonetics*. The MIT Press

Sun, C. (2006). *Chinese: A Linguistic Introduction*. Cambridge University Press.

Teixeria, J. & Fernandes, P. (2015). Acoustic Analysis of Vocal Dysphonia. *Procedia Computer Science* 64, 466-473.

Urua, E. & Udoh, E. (2017). The prosodic status of glides in Anaañ reduplication. In Lindsey, G. & Nevins, A. (Eds., Vol. 14, pp. 297-319) *Sonic Signatures*. John Benjamins Publishing Company.

Villafaña-Dalcher C., Knight, A., Jones, M. (2008). Cue switching in the perception of approximants: Evidence from two English dialects. *University of Pennsylvania Working Papers in Linguistics* 14(2, Article 9), 63-71.

Wang, J. (王佳林). (2011). A sociolinguistic study on the initial [ʋ]-ization of the Harbin dialect (哈尔滨话合口呼零声母[ʋ]化的社会语言学研究). MA Thesis, Heilongjian University. Accessed via Baidu (百度).

Wiener, S. & Shih, Y. (2011). Divergent places of articulation: [w] and [ʋ] in modern spoken Mandarin. In Jing-Schmidt, Zhuo (Ed.) *Proceedings of the 23rd North American Conference on Chinese Linguistics* 1, 173-190.

Ying, J (英君). (2011). Standards and Variation in Pǔtōnghuà Pronunciation (普通话语音的规范与变异). *Journal of Yangtze University* (*长江大学学报，社会科学版*) 34(8), 76-77. Accessed via Baidu (百度).

Zamponi, V. (2021). Effect of sex hormones on human voice physiology: from childhood to senescence. *Hormones* 20, 691-696. Springer.

Zhou, J. (周锦国). (2003). A Phenomenon of Phonetic Change in Pǔtōnghuà: Xinven vs. Xinwen (News) in the TV show of “New 30” (现代汉语普通话语音中的一种音变现象–从‘新闻 30 分’重的新闻（xinven）谈起). *Applied Linguistics* (*语言文字应用*) 1, 116-119. Accessed via Baidu (百度).

Zhou, Y. (周有光) (2009). The historical evolution of Chinese languages and scripts (中国语文的时代演进). People’s Literature Publishing House (人民文学出版社).

(1992). *Ethnologue: Languages of the World*. Grimes, B. F. (Ed.). Print.

Appendix A Code

<u>Praat Script:</u>

```
# This script is based off of one originally written by
# Shigeto Kawahara (author: April 2010) and Joseph LeBeau (mod: Oct. 2013)
# Final version by Jordan Crowley, 23 May, 2022
#
# This version finds matching pairs of .wav and .TextGrids files in a
selected directory
# a selected directory, then extracts from each non-empty interval:
# label, duration, lowest average F2, F2 slope, and mean HNR
# This was intended to be used only on [w] onsets on the
male/female/north/south divide

# There is some error with the ceiling parameter, so be sure to change
between 5000/5500 manually

form
     sentence Directory C:\
     choice Gender 1
          button male
          button female
     choice Region 1
          button north
          button south
endform

Create Strings as file list... wavlist 'directory$'/*.wav
number_files = Get number of strings

if gender$ = "male" and region$ = "north"
     ceiling = 5000 and gender$ = "male" and region$ = "north"

elsif gender$ = "male" and region$ = "south"
     ceiling = 5000 and gender$ = "male" and region$ = "south"

elsif gender$ = "female" and region$ = "north"
     ceiling = 5500 and gender$ = "female" and region$ = "north"

elsif gender$ = "female" and region$ = "south"
     ceiling = 5500 and gender$ = "female" and region$ = "south"
endif

appendInfoLine: "speaker ", "filename ", "gender ", "region ", "word ",
"duration ", "F2min ", "F2max ", "meanF2 ", "slope ", "HNR "

for i from 1 to number_files
     select Strings wavlist
     filename$ = Get string... i
     Read from file... 'directory$'/'filename$'
     soundname$ = selected$ ("Sound")
     formants = To Formant (burg): 0, 5, 5500, 0.025, 50
     Read from file... 'directory$'/'soundname$'.TextGrid
     gridname$ = selected$("TextGrid")
     plus Sound 'soundname$'
     Extract non-empty intervals... 1 Preserve times
     hnr = To Harmonicity (cc)... 0.01 70 0.1 1
     select TextGrid 'gridname$'
     if filename$ <> ""
          speaker$ = left$ (filename$, (length (gridname$) - 4))
     endif

     n = Get number of intervals: 1
     for k from 1 to n
          label$ = Get label of interval: 1, k
          if label$ <> ""
               t1 = Get start time of interval: 1, k
               t2 = Get end time of interval: 1, k
               duration = t2 - t1
               select formants
               f2min = Get value at time: 2, t1, "hertz", "linear"
               f2max = Get value at time: 2, t2, "hertz", "linear"
               mean = (f2min + f2max)/2
               slope = ((f2max - f2min)/duration)/1000
               select hnr
               meanHNR = Get mean... 0 0
          meansdHNR = Get standard deviation... 0 0
               select TextGrid 'gridname$'
               appendInfoLine: speaker$, " ", gridname$, " ", gender$, " ",
region$, " ", label$, " ", 'duration:3'*1000, " ", 'f2min:1', " ",
'f2max:1', " ", 'mean:1', " ", 'slope:4', " ", 'meanHNR:3'
               removeObject: formants
          endif
     endfor
     select all
     minus Strings wavlist
          Remove
endfor

select all
Remove
```

R Code:

~~~
# Table

Data Frame
```{r}
table = read.delim ("data/thesis.txt", stringsAsFactors=TRUE)
table
levels (table$speaker)
mean.table = aggregate (x = table[c("f2min", "slope", "duration", "hnr")], by = table[c("gender", "region",
"word")], FUN = mean, na.rm = TRUE)
sd.table = aggregate (x = table[c("f2min", "slope", "duration", "hnr")], by = table[c("gender", "region",
"word")], FUN = sd, na.rm = TRUE)
mean.table
sd.table
r="Region Contrast"
g="Gender Contrast"
w="Word Contrast"
f="F2 Frequency (Hz)"
h="HNR (dB)"
s="Slope (Hz/ms)"
```
~~~

~~~
# First Set of Contrasts, Models and Graphs
```{r}
gender.contrast = cbind (c(+0.5, -0.5)) #Literature suggests males will have lower F2
colnames (gender.contrast) = c("-M+F")  #male coded as left, female coded as right
contrasts (table$gender) = gender.contrast
contrasts (table$gender)
levels (table$gender)

region.contrast = cbind (c(+0.5, -0.5)) #Literature suggests south will have lower F2
colnames (region.contrast) = c("+N-S")  #south coded as left, north coded as right
contrasts (table$region) = region.contrast
contrasts(table$region)
levels (table$region)

word.contrast = cbind (c(-0.5, +0.5))     #Literature suggests "wei" will have lower F2
colnames (word.contrast) = c("-WEI+WEN")  #"wei" coded as left, "wen" coded as right
contrasts (table$word) = word.contrast
contrasts(table$word)
levels (table$word)

low_model = lmerTest::lmer (formula = f2min ~ region * gender * word + (1 + word| speaker), data = table,
REML = TRUE)
summary(low_model)
#confint(low_model) #Produces a weird error if you uncomment, but still produces a confint at the end.

gender.table = read.delim ("data/thesis_gender.txt", stringsAsFactors=TRUE)
gender.table$gender.symm.MF = c (rep (+0.4, 710), rep (-0.4, 290))   # females right coded: in table, female
first
plot (gender.table$gender.symm.MF, gender.table$f2min, pch='__', xlab=g, ylab=f)
abline(lm(gender.table$f2min ~ gender.table$gender.symm.MF))

region.table = read.delim("data/thesis_region.txt", stringsAsFactors=TRUE)
region.table$region.symm.SN = c (rep (+0.4, 650), rep (-0.4, 350)) #south left coded: in table, north first
plot (region.table$region.symm.SN, region.table$f2min, pch='__', xlab=r, ylab=f)
abline(lm(region.table$f2min ~ region.table$region.symm.SN))

word.table = read.delim("data/thesis_word.txt", stringsAsFactors=TRUE)
word.table$word.symm.WEIWEN = c (rep (-0.4, 500), rep (+0.4, 500)) #WEI left coded: in table, WEI first
plot (word.table$word.symm.WEIWEN, region.table$f2min, pch="__", xlab=w, ylab=f)
abline(lm(word.table$f2min ~ word.table$word.symm.WEIWEN))
```

# Second Set of Contrasts, Models, and Graphs

```{r}
gender.contrast = cbind (c(-0.5, +0.5)) #Literature suggests females will have flatter slope and lower HNR
colnames (gender.contrast) = c("-F+M")  #female coded left
contrasts (table$gender) = gender.contrast
contrasts (table$gender)

region.contrast = cbind (c(-0.5, +0.5)) #Literature suggests north will have flatter slope and lower HNR
colnames (region.contrast) = c("-N+S")  #north coded left
contrasts (table$region) = region.contrast
contrasts(table$region)

word.contrast = cbind (c(+0.5, -0.5))     #Literature suggests "WEN" will have flatter slope and lower HNR
colnames (word.contrast) = c("-WEN+WEI")  #"WEN" coded left
contrasts (table$word) = word.contrast
contrasts(table$word)
~~~

```
slope_model = lmerTest::lmer (formula = slope ~ region * gender * word + (1 + word| speaker), data = table,
REML = TRUE)
summary(slope_model)
#confint(slope_model)

gender.table$gender.symm.FM = c (rep (-0.4, 710), rep (+0.4, 290))   # females left coded: in table, female first
plot (gender.table$gender.symm.FM, gender.table$slope, pch='__', xlab=g, ylab=s)
abline(lm(gender.table$slope ~ gender.table$gender.symm.FM))

region.table$region.symm.NS = c (rep (-0.4, 650), rep (+0.4, 350)) #North left coded: in table, north first
plot (region.table$region.symm.NS, region.table$slope, pch='__', xlab=r, ylab=s)
abline(lm(region.table$slope ~ region.table$region.symm.NS))

word.table$word.symm.WENWEI = c (rep (+0.4, 500), rep (-0.4, 500)) #WEN left coded: in table, WEI first
plot (word.table$word.symm.WENWEI, region.table$slope, pch="__", xlab=w, ylab=s)
abline(lm(word.table$slope ~ word.table$word.symm.WENWEI))

hnr_model = lmerTest::lmer (formula = hnr ~ region * gender * word + (1 + word| speaker), data = table,
REML = TRUE)
summary(hnr_model)
#confint(hnr_model)

plot (gender.table$gender.symm.FM, gender.table$hnr, pch='__', xlab=g, ylab=h)
abline(lm(gender.table$hnr ~ gender.table$gender.symm.FM))

plot (region.table$region.symm.NS, region.table$hnr, pch='__', xlab=r, ylab=h)
abline(lm(region.table$hnr ~ region.table$region.symm.NS))

plot (word.table$word.symm.WENWEI, region.table$hnr, pch="__", xlab=w, ylab=h)
abline(lm(word.table$hnr ~ word.table$word.symm.WENWEI))

boxplot(table$hnr ~ (table$word + table$gender), outline=FALSE, xlab="Word and Gender", ylab="HNR
(dB)",
+         main="HNR Interaction Effects")
```

Appendix B Model Summaries

F2$_{\text{min}}$ Model:

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method
['lmerModLmerTest']
Formula: f2min ~ region * gender * word + (1 + word | speaker)
   Data: table

REML criterion at convergence: 13510.5

Scaled residuals:
    Min      1Q  Median      3Q     Max 
-2.6311 -0.5817 -0.0653  0.4701  5.9922 

Random effects:
 Groups   Name         Variance Std.Dev. Corr 
 speaker  (Intercept)  17664    132.91        
          word-WEI+WEN  5463     73.91   -0.35
 Residual              38161    195.35        
Number of obs: 1000, groups:  speaker, 100

Fixed effects:
                                   Estimate Std. Error      df t value Pr(>|t|)    
(Intercept)                         1125.54      16.96   96.00  66.346  < 2e-16 ***
region+N-S                           127.95      33.93   96.00   3.771 0.000281 ***
gender-M+F                           157.28      33.93   96.00   4.636 1.12e-05 ***
word-WEI+WEN                         -59.57      16.66   96.00  -3.575 0.000551 ***
region+N-S:gender-M+F                -79.00      67.86   96.00  -1.164 0.247246    
region+N-S:word-WEI+WEN               25.96      33.33   96.00   0.779 0.437886    
gender-M+F:word-WEI+WEN              -17.76      33.33   96.00  -0.533 0.595288    
region+N-S:gender-M+F:word-WEI+WEN  -102.96      66.66   96.00  -1.545 0.125741    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) rg+N-S gn-M+F w-WEI+ rg+N-S:-M+F r+N-S:-W g-M+F:
region+N-S  -0.306                                                 
gender-M+F  -0.424  0.136                                          
wrd-WEI+WEN -0.162  0.050  0.069                                   
rg+N-S:-M+F  0.136 -0.424 -0.306 -0.022                            
r+N-S:-WEI+  0.050 -0.162 -0.022 -0.306  0.069                     
g-M+F:-WEI+  0.069 -0.022 -0.162 -0.424  0.050       0.136         
r+N-S:-M+F: -0.022  0.069  0.050  0.136 -0.162      -0.424   -0.306
```

F2$_{\text{min}}$ Confidence Interval:

```
                                        2.5 %       97.5 %
(Intercept)                        1092.64541 1158.43285162
region+N-S                           62.16162  193.73650324
gender-M+F                           91.49471  223.06959798
word-WEI+WEN                        -91.88400  -27.25986127
region+N-S:gender-M+F              -210.57297   52.57679596
region+N-S:word-WEI+WEN             -38.66000   90.58827349
gender-M+F:word-WEI+WEN             -82.38758   46.86069849
region+N-S:gender-M+F:word-WEI+WEN -232.20716   26.28938913
```

F2 Slope Model:

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method
['lmerModLmerTest']
Formula: slope ~ region * gender * word + (1 + word | speaker)
   Data: table

REML criterion at convergence: 5651.2

Scaled residuals: 
    Min      1Q  Median      3Q     Max 
-3.4263 -0.5323 -0.0206  0.5458  4.3885 

Random effects:
 Groups   Name         Variance Std.Dev. Corr
 speaker  (Intercept)   4.043   2.011        
          word-WEN+WEI  3.538   1.881    0.02
 Residual              14.051   3.749        
Number of obs: 1000, groups:  speaker, 100

Fixed effects:
                                   Estimate Std. Error       df t value Pr(>|t|)
(Intercept)                         8.99588    0.27019 95.99923  33.295  < 2e-16
***
region-N+S                          0.77348    0.54037 95.99923   1.431    0.156
gender-F+M                         -2.23124    0.54037 95.99923  -4.129 7.77e-05
***
word-WEN+WEI                        3.50653    0.35030 96.00022  10.010  < 2e-16
***
region-N+S:gender-F+M              -0.35704    1.08075 95.99923  -0.330    0.742
region-N+S:word-WEN+WEI            -0.50348    0.70060 96.00022  -0.719    0.474
gender-F+M:word-WEN+WEI             0.07193    0.70060 96.00022   0.103    0.918
region-N+S:gender-F+M:word-WEN+WEI -0.04099    1.40120 96.00022  -0.029    0.977
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) rg-N+S gn-F+M w-WEN+ rg-N+S:-F+M r-N+S:-W g-F+M:
region-N+S  0.306                                                  
gender-F+M  0.424  0.136                                           
wrd-WEN+WEI 0.013  0.004  0.005                                    
rg-N+S:-F+M 0.136  0.424  0.306  0.002                             
r-N+S:-WEN+ 0.004  0.013  0.002  0.306  0.005                      
g-F+M:-WEN+ 0.005  0.002  0.013  0.424  0.004       0.136          
r-N+S:-F+M: 0.002  0.005  0.004  0.136  0.013       0.424    0.306 
```

F2 Slope Confidence Interval:

```
                                        2.5 %     97.5 %
(Intercept)                         8.4720018  9.5197588
region-N+S                         -0.2742724  1.8212416
gender-F+M                         -3.2789998 -1.1834858
word-WEN+WEI                        2.8273116  4.1857469
region-N+S:gender-F+M              -2.4525508  1.7384773
region-N+S:word-WEN+WEI            -1.8619175  0.8549535
gender-F+M:word-WEN+WEI            -1.2865017  1.4303690
region-N+S:gender-F+M:word-WEN+WEI -2.7578596  2.6758817
```

HNR Model:

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method
['lmerModLmerTest']
Formula: hnr ~ region * gender * word + (1 + word | speaker)
   Data: table

REML criterion at convergence: 5594.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.7587 -0.5554 -0.0034  0.5976  3.1551

Random effects:
 Groups   Name         Variance Std.Dev. Corr
 speaker  (Intercept)   4.824   2.1965
          word-WEN+WEI  0.240   0.4899   -1.00
 Residual              13.696   3.7008
Number of obs: 1000, groups:  speaker, 100

Fixed effects:
                                   Estimate Std. Error       df t value Pr(>|t|)
(Intercept)                         12.6370     0.2881  96.0047  43.867  < 2e-16
***
region-N+S                          -0.1523     0.5762  96.0047  -0.264   0.7921
gender-F+M                          -2.7801     0.5762  96.0047  -4.825 5.26e-06
***
word-WEN+WEI                         0.6204     0.2768 480.6466   2.241   0.0255 *
region-N+S:gender-F+M                0.4727     1.1523  96.0047   0.410   0.6825
region-N+S:word-WEN+WEI              0.4282     0.5536 480.6466   0.774   0.4396
gender-F+M:word-WEN+WEI              1.1116     0.5536 480.6466   2.008   0.0452 *
region-N+S:gender-F+M:word-WEN+WEI  -0.8864     1.1072 480.6466  -0.801   0.4238
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) rg-N+S gn-F+M w-WEN+ rg-N+S:-F+M r-N+S:-W g-F+M:
region-N+S   0.306
gender-F+M   0.424  0.136
wrd-WEN+WEI -0.181 -0.055 -0.077
rg-N+S:-F+M  0.136  0.424  0.306 -0.025
r-N+S:-WEN+ -0.055 -0.181 -0.025  0.306 -0.077
g-F+M:-WEN+ -0.077 -0.025 -0.181  0.424 -0.055       0.136
r-N+S:-F+M: -0.025 -0.077 -0.055  0.136 -0.181       0.424    0.306
```

HNR Confidence Interval:

```
                                         2.5 %     97.5 %
(Intercept)                        12.07840089 13.1955969
region-N+S                         -1.26950189  0.9648902
gender-F+M                         -3.89729283 -1.6629007
word-WEN+WEI                        0.07855266  1.1622265
region-N+S:gender-F+M              -1.76166255  2.7071217
region-N+S:word-WEN+WEI            -0.65546905  1.5118787
gender-F+M:word-WEN+WEI             0.02787958  2.1952273
region-N+S:gender-F+M:word-WEN+WEI -3.05370263  1.2809928
```