UNIVERSITY OF AMSTERDAM
Amsterdam Center for Language and Communication

Vowels in Shanghai Chinese: Acoustic realization and modeling of their diachronic change with an LSTM network

ResMA Thesis Linguistics and Communication

Chengjia Ye (Student ID: 13529374)

Supervisor: Prof. dr. Paul Boersma

Second Examiner: Dr. Silke Hamann

Graduate School of Humanities, the University of Amsterdam

Amsterdam, the Netherlands

June 2023

Abstract

Shanghai Chinese changed drastically in the past decades. This study provides the most state-of-the-art documentation of the acoustic realization of its vowel system in the generation born after 1995 and models the diachronic vowel change using a Long-Short Term Memory (LSTM) network with data obtained from a monosyllabic wordlist-reading task. The effects of lexical tones (syllable structure and tone register) were under investigation. A new monophthong system is proposed with minor adjustments on transcription, including 10 open-ended vowels /i, y, e, ɛ, ø, ɤ, ɑ, ɯ, u, ɔ/ that are associated with unchecked tones (T1, T2 and T3) and 5 glottal-ended vowels /ɪʔ, ʏʔ, əʔ, ɐʔ, oʔ/ that are associated with checked tones (T4 and T5). More than two-thirds of the inventory gather in the upper half of the oral cavity, and the feature of roundness is common with five round-unround pairs.

Statistical analysis on the acoustic data of 23 young generation participants born after 1995 shows that the former are around 2.4 times longer than the latter when produced in isolation. Duration of vowels according to lexical tones is ranked in a descending order as: T3 ≈ T2 > T5 > T4 (in which T3 and T5 are low-register tones with a phonologically voiced onset). The vowel space of open-ended vowels is more peripheral or extensive while that of glottal-ended ones is more central. Apart from syllable structure, tone register seems to play a limited role in deciding vowel height, while gender plays an important role in vowel backness. The breathiness which used to occur at the beginning part of low-register vowels in the old generation was not detected in the current study. This might be related to the fact that the breathiness is fading out in speech production of the younger generation under strong influence of Mandarin Chinese. However, open-ended vowels seem to have a higher spectral tilt than glottal-ended ones. We also found out that low-register tone T3 is accompanied with creaky voice, which is probably a result of its low *f*0.

A four-layer LSTM network with duration and vowel space (F1 and F2) measures as features was built with data from the 23 young generation participants and 6 elder participants aged between 40 and 60 obtained an F-score of 0.60 with a well-balanced performance between Type I and Type II errors. Error analysis revealed three reasons behind the wrong predictions: (1) The vowels are prohibited from existing in the same context by phonotactic rules, albeit their overlapping space; (2) The vowels are in process of convergence or divergence under the impact of Mandarin Chinese; and (3) Some features like F3 are not included in the model.

Combing the findings, we get a clear picture of the rich vowel system in Shanghai Chinese within which most vowels are distinctive either phonetically in terms of duration and/or vowel space or at a phonological level as conditioned by phonotactic rules. Meanwhile, some vowels are under dynamic diachronic change.

Key words: Shanghai Chinese, phonetics, phonology, tone, phonation type, vowel duration, Long-Short Term Memory (LSTM)

1. Introduction

Shanghai Chinese, also known as Shanghainese or Shanghai Wu, is a representative subvariant of Wu Chinese (ISO 639-3 code: wuu) that is mainly spoken in the urban and surrounding area in the city of Shanghai. It is considered as an underdocumented language despite a number of comprehensive studies on its sound system (e.g., Edkins, 1853; Chao, 1928; Xu & Tang, 1988; Chen & Gussenhoven, 2015) and studies centering on the phonetic properties of its consonants (e.g., Ren, 1992), vowels (e.g., Chen, 2008), lexical tones and tone sandhi (e.g., Zee & Maddison, 1980; Zhang & Meng, 2016; Ling & Liang, 2019). As noted in Xu & Tang (1988), its phonological system has undergone drastic changes in the past few decades. One cause of the huge historical change is the influence of Mandarin Chinese (Gao et al., 2019; Chen & Gussenhoven, 2015).

The current study aims to investigate the status quo and the diachronic change of acoustic realization of Shanghai Chinese vowels. By doing that, we may gain insight into the latest development of the vowel system in Shanghai Chinese. More importantly, it may also reveal the general trend in vowel inventories under the impact of language contact.

1. 1 Diachronic change in Shanghai Chinese

Shanghai Chinese has changed greatly at both the phonological and phonetic level. As a tonal language, it currently has five distinctive lexical tones (Xu & Tang, 1988; Qian, 2003) evolving from the eight tones reported by Edkins (1853) in accordance with the Middle Chinese tone system. In Middle Chinese, there are two registers (namely High and Low) of tones, and four contour patterns in each register, i.e., level, rising, departing (falling) and entering (checked), yielding eight lexical tones. As early as Edkin's (1853) time, the low-rising tone began to diverge and merge with both the low-level and low-departing tone. In the generation born in the old urban area before the 1930s, the low-level tone also merged with the low-departing tone (Xu & Tang, 1988), reducing the number of tones to six. In addition, the high-rising tone started the merger with the high-departing tone around the same time. Thus, there are three tones remaining in the high register in Shanghai Chinese, i.e., T1 (high-level), T2 (high-departing) and T4 (checked short-high) and two in the low register, i.e., T3 (low-departing) and T5 (checked short-rising) today. It is noteworthy that their actual pitch values and contour patterns no longer match the names after the systematic change. T1, in spite of being categorized as a high-level tone, currently has a high-falling contour pattern. The contour patterns of T2 and T3, the two departing tones, resemble each other with a slight rise towards the end.

The diachronic change in lexical tones is driven by the strong impact of Mandarin Chinese (Z. Chen, 2007) as well as the internal factors within Shanghai Chinese. In terms of the merger of the three unchecked lexical tones in the low register, Z. Chen (2007) noted that the actual pitch values and contour patterns of these tones were similar, and that the pitch value of many words with a low-departing tone was unstable, sometimes transcribed as a low-level tone in Edkins (1853), providing preconditions for merging. In addition, the role of tones is limited in utterances because of tone sandhi (see Zee & Maddieson, 1980; Duanmu, 1999). Only the initial syllable in a prosodic unit retains its pitch contour pattern, while that of all other syllables

gets neutralized and realized as the relative pitch height (i.e., being High or Low), making Shanghai Chinese similar to pitch-accent languages like Japanese.

Furthermore, the phonological system of Shanghai Chinese is highly integrated in the sense that the phonotactic rules contain repetitive information from various perspectives that may seem redundant and allows simplification. The tone register and the voicing of syllable onset are co-conditioned by each other. Syllables with a voiced onset only cooccur with low-register tones, i.e., T3 or T5, whilst voiceless onsets only cooccur with high-register tones, i.e., T1, T2 or T4. The pitch contour patterns are associated with syllable structures. The first three unchecked contour patterns, i.e., level, rising and departing, are associated with open syllables or nasal-ended syllables that end with /ŋ/, a remnant from Middle Chinese that is usually realized as the nasalization of the preceding vowel (Svantesson, 1989). The checked tones, on the contrary, are only associated with syllables ending with a glottal stop /ʔ/, corresponding to the three stop codas /p, t, k/ in Middle Chinese. Table 1 shows the co-occurrence of the five existing lexical tones, onsets and syllable structures in Shanghai Chinese.

Table 1. The co-occurrence of tone, onset, and syllable structure in Shanghai Chinese

register	syllable onset	level	rising	departing	entering
			/(C)V/ or /(C)Vŋ/		/(C)Vʔ/
High	[-voice]	T1	--	T2	T4
Low	[+voice]	--	--	T3	T5

The phonetic implementation of consonantal onsets changed considerably, as in the voicing of obstruents. At the phonemic level, there is a three-way laryngeal contrast among voiceless unaspirated, voiceless aspirated, and voiced obstruents as a remnant of Middle Chinese. Today, voiced obstruents, however, are distinct from voiceless unaspirated ones in terms of voice onset time (VOT) only at the non-initial position of a prosodic unit (Ren, 1992; Chen, 2011). In initial position, the voiced obstruents are often realized with a shorter closure duration than their voiceless unaspirated counterparts (Shen et al., 1987) and slight breathiness (with a higher H1–H2 difference, see Chai & Garellek, 2022 for a review on this measure) at the release of constriction and the beginning of the following vowel (Cao & Maddieson, 1992; Gao & Hallé, 2017). Recent studies also found that the breathy phonation embedded in low-register tones is gradually fading away among the younger generation with no systematical voicing difference detected between the two registers in syllables produced by the younger generation born in the late 1980s and the early 1990s (Gao, 2016; Gao & Hallé, 2017). The loss of breathiness in production is also seen as a consequence of Mandarin Chinese influence (Gao et al., 2019).

The number of rhymes also decreased from 51 to 32 within three generations (see Table 2, summarized from Xu & Tang, 1988). Several trends of vowel shift can be observed. In terms of open syllables, from the old system to the middle system, /ɥ/ merged with /ʐ̩/, /e/ merged with /ɛ/ in both monophthong and diphthong, while a new diphthong /uø/ emerged. From the middle system to the new system, /uø/ and /yø/ merged with /ø/ with the loss of the first part of the diphthong. Some words with /yø/, at the same time, merged with /y/ with the loss of the second part. In terms of glottal-ended syllables, the reduction was even more radical. Rhymes

beginning with /i/ and /u/ were simplified and only one of each remains. /œʔ/ no longer distinguished from /oʔ/ or /əʔ/, while /əʔ/ further merged with /oʔ/. /oʔ/ also admitted /ɔʔ/ at the middle system. Apart from these trends, Chen (2008) transcribed /iɪʔ/ and /ʏɪʔ/ as /ɪʔ/ and /ʏʔ/, which was also a minor tendency mentioned in Xu & Tang (1988). They are therefore treated as monophthongs with a glottal coda in this study.

Table 2. The rhymes across three generations in Shanghai Chinese (c.f., Xu & Tang, 1988)

	Old (born before 1930)	**Middle** (born between 1930-1965)	**New** (born after 1965)
open syllable	ʐ̩[1] ɥ ɑ ɔ o ɤ e ɛ ø	ʐ̩ ʌ[2] ɔ o ɤ ɛ ø	ʐ̩ ʌ ɔ o ɤ ɛ ø
	i iɑ[3] iɔ iɤ iɛ iɪ iu	i iʌ iɔ iɤ iɛ	i iʌ iɔ iɤ
	u uɑ ue uɛ	u uʌ uɛ uø	u uʌ uɛ
	y yø	y yø	y
nasal ended	ã ɑ̃ əŋ oŋ	ã ɑ̃ ən oŋ	ʌ̃[2] ən oŋ
	iã iɑ̃ iəŋ ioŋ	iã iɑ̃ in ioŋ	iʌ̃ in ioŋ
	uã uɑ̃ uəŋ	uã uɑ̃ uən	uʌ̃ uən
		yn	yn
glottal ended	ʌʔ œʔ oʔ ɔʔ əʔ	ʌʔ oʔ əʔ	ɐʔ oʔ
	iʌʔ ioʔ iɪʔ iəʔ	iʌʔ ioʔ iɪʔ	iɪʔ
	uʌʔ uœʔ uoʔ uəʔ	uʌʔ uəʔ	uɐʔ
	ʏœʔ	ʏɪʔ	ʏɪʔ
others	əl m̩ n̩ ŋ̩	əl m̩ n̩ ŋ̩	əl m̩ n̩
Total	51	43	32

Note. [1] /ʐ̩/ is an allophone of /i/ (Svantesson, 1989) when preceded by sibilants /s, z, ts, tsʰ/ that is traditionally transcribed as /ɿ/ in Xu & Tang (1988). It is often seen as a syllabic fricative, whereas its vowel spectral qualities and degree of friction vary (Chen & Gussenhoven, 2015). [2] /ʌ/ represents the low central vowel between the front /a/ and back /ɑ/ which also exists in Mandarin Chinese but is absent on the IPA chart. It is transcribed as /ɐ/ in Chen (2008). However, according to Xu & Tang (1988), /ɐʔ/ is the result of the merger of /ʌʔ/ and /əʔ/. [3] Diphthongs beginning with /i/ or /u/ are analyzed as glide (either /j/ or /w/) + monophthongs in Chen & Gussenhoven (2015).

1. 2 Shanghai Chinese vowels

In an earlier work on the phonetic realization of vowels in Shanghai Chinese, Svantesson (1989) analyzed the frequencies of the first three formants embedded in a carrier sentence produced by three native speakers born between 1937 and 1964 and found no significant influence of tones and phonation types on the vowel production. However, as only one value for each formant was obtained at a single trial, the variation pattern within a syllable could not be reflected. Thereafter, Chen (2008) conducted a more comprehensive production study with 13 native speaking participants born between 1935 and 1950. She adopted two complementary sets for vowels embedded in open syllables /a, ɔ, o, ɤ, ɛ, ø, i, u, y/ and closed syllables (either with a nasal or a glottal coda) /ɐ, ʊ, ə, ɪ, ʏ/, which is slightly different from the system introduced by Xu & Tang (1988) and that of Svantesson (1989). These two vowel sets

demonstrated significant difference in terms of duration (open vs. glottal ended) and vowel space. The vowels in open syllables were much longer and were distributed more extensively on the vowel space than those in glottal-ended syllables. Additionally, the effect of consonantal onset and prosodic prominence was confirmed but turned out to be different for these two sets of vowels.

So far, the few studies on the vowel space in Shanghai Chinese did not include data across generations or data on the younger generation (especially those born after 1995), hence the lack of documentation on the latest acoustic information of vowels and on the tendency for vowel shift in Shanghai Chinese.

On the other hand, some factors other than onset and prosodic prominence also have an effect on the vowel, especially in the case of Shanghai Chinese in which vowels are associated with tone contour patterns and syllable structures. Tone registers, for instance, bring about a contrast between breathy and modal phonation. As mentioned in Section 1.1, the beginning part of vowels with a low-register tone (T3 or T5) is often realized with breathy phonation, shown as a higher H1–H2 difference (Cao & Maddieson, 1992) than that of vowels associated with a high-register tone (T1, T2 or T4). More specifically, when it occurs, it is the most salient at the vowel onset and is maintained till around the midpoint of the vowel (Gao & Hallé, 2017) although the phonation contrast between the two registers becomes less systematic and less common among young native speakers (Tian & Kuang, 2016). It remains unclear whether the change in voice quality is also related to phonetic or phonological factors.

Nonetheless, the low tones in other variants of Chinese like Mandarin (e.g., Kuang, 2017; Huang, 2020) and Cantonese (e.g., Yu & Lam, 2014; Zhang & Kirby, 2020) as well as other tonal languages like Vietnamese (Brunelle, 2009) may occasionally carry a creaky phonation, which serves as an important perceptual cue in identifying tones whose pitch contour patterns closely resemble each other. The creakiness in these languages is systematically tied to *f*0 in perception, and more particularly, to the extra-low *f*0 (Huang, 2020). It is therefore plausible to hypothesize that the low-register tones (T3 and T5) may also have some creakiness in production, as the pitch contour patterns in Shanghai Chinese lexical tones do not differ much between the two registers. Since creaky phonation is opposite to breathy phonation, it may explain the disappearance of breathy phonation in the middle of the vowel and the loss of breathiness in the younger generation.

1. 3 Goals of the present study

This paper has two goals. The primary goal is to investigate the current acoustic realization of the 10 monophthongs in open syllables and their 5 counterparts with a glottal stop in the end that are documented in the phoneme inventory of Shanghai Chinese used by the generation born between 1930 and 1965 (Xu & Tang, 1988) among the younger generation born after 1995. However, /ʐ̩/ is excluded from the monophthongs due to its various realization of vowel qualities and frication, while /e/ from the inventory used by elderly people born before the 1930s is included as it has been reported to diverge again from /ɛ/ in the younger generation (Gu, 2004) as the result of Mandarin Chinese influence. Table 3 summarizes the 15 monophthongs to be examined in this study.

Table 3. The 15 Shanghai Chinese monophthongs to be examined in the study

	front					middle		back				
open ended	i	y	e	ø	ɛ			u	o	ɤ	ɔ	ɑ
glottal ended	ɪʔ	ʏʔ				əʔ	ɐʔ		oʔ			

By testing the acoustic features of these 15 vowels in monosyllabic words produced in isolation, we may summarize the tendencies of vowel shift and provide the latest acoustic information about the vowel inventory in Shanghai Chinese. Attention is largely paid to vowel duration, vowel space and phonation type. Chen's (2008) study showed that vowels in an open syllable are on average 1.5 times longer than those in a glottal-ended syllable associated with a checked tone when embedded in a carrier sentence. This duration contrast is expected to be more salient for words in isolation. At the same time, we attempt to understand whether lexical tone contour patterns have an effect on the acoustic realizations of vowels. As prosodic prominence in general enhances the F2 distinction between front and back vowels and increases the F1 value (Chen, 2008), while tone register and onset voicing may affect the phonation type of the first half of a vowel (Cao & Maddieson, 1992; Gao & Hallé, 2017), we intend to see whether there is also a difference in formant frequencies between high tones (T2 and T4) and low tones (T3 and T5). T1 (high-level) is not considered in this study for the absence of its counterpart in the low register.

In terms of voice quality, we examine two seemingly contradictory non-modal phonation types that may appear in low tones with the following two research questions: (1) Whether the breathiness (realized as a higher H1–H2 difference) can still be detected at the beginning part of vowels with low-register tones in the younger generation; (2) Whether the low-register tones are sometimes produced with a creaky phonation type, characterized as having extra-low *f*0, which is the primary cue in other Chinese variants, and/or irregular *f*0. Although H1–H2 difference may also reflect the contrast between modal and creaky phonation, it has many limitations and is not seen as an important cue in this scenario (Chai & Garellek, 2022).

The second goal of this study is to model the tendency of Shanghai Chinese vowel shift using a semi-supervised Long-Short Term Memory (LSTM) network, a recurrent neural network, which also provides a real-life case for modeling language change. Neural networks, as popular non-linear machine learning models, are widely used in practice for speech recognition and in theory to simulate phonetic and phonological phenomena such as speech acquisition and formation of phonemic categories (e.g., Beguš, 2020; Boersma et al., 2020, 2022). To our knowledge, there are few neural network models modeling diachronic change in a language, which is an important aspect of a language model (Boersma et al., 2020). This is partially due to the difficulty in getting real-life data on language change as it usually happens within a wide time span. Hence the drastic change in Shanghai Chinese vowels may serve as a good material. An LSTM is ideal for modeling language change as it can capture the temporal dependencies, retain memories from both decades ago and the recent past while weighing their connections and biases differently, and resist the interference of noise and missing values that are inevitable in historical linguistic data.

2. Methods

2. 1 Participants

23 Shanghai Chinese native speakers born after 1995 (12 females and 11 males, aged from 16 to 27 years at the time of recording, *mean* = 22.6, *SD* = 3.14) participated in the wordlist reading task. The participants are from the urban or close-by (3 participants) suburban area of Shanghai city. Apart from Shanghai Chinese, they all have learnt Mandarin Chinese before the age of seven and have learnt English as a foreign language at school as part of the compulsory education. Another 6 elder native speakers (4 females and 2 males, aged from 40 to 60 years, *mean* = 49.67, *SD* = 7.34) were recruited as the control group for the modeling of vowel shift. The elder participants have also learnt at school or self-learnt Mandarin Chinese apart from Shanghai Chinese. All 29 participants self-reported to have normal hearing and no speech impairments and to be actively speaking and passively listening to Shanghai Chinese in their daily life.

All of them participated in this task on a voluntary basis with consent obtained prior to the test. The experiment had been approved by the Ethics Committee of Amsterdam Institute for Humanities Research at the University of Amsterdam.

2. 2 Materials

A wordlist reading task was designed to test the acoustic realization of the 15 vowels from two complementary sets in Shanghai Chinese. All stimuli are mono-syllabic words so as to rule out the effect of tone sandhi on tone contour patterns and the effect of prosodic prominence on vowel space. The latter has been proved by Chen's (2008) experiment with carrier sentences in which the lexical tones of the target words belong to the low register, i.e., either T3 for open syllables or T5 for glottal-ended syllables. To explore the effect of tone registers on the acoustic realization of vowels, the tone is manipulated between T2/T4 (high tones) and T3/T5 (low tones).

We chose /d/ as the onset for T3 and T5, and its voiceless unaspirated counterpart /t/ for T2 and T4 because voiceless onsets may only occur in syllables whose tone falls into the high register (i.e., T1, T2 and T4) while voiced onsets may only occur in syllables whose tone falls into the low register (i.e., T3 and T5). Voiceless aspirated onsets are not taken into consideration as they can be easily distinguished from the other two types of onsets for their high positive VOT values, except for /tʰɑ/ to avoid the potential ambiguity engendered by the other common pronunciations that the only two frequently used words associated with /tɑ/ carry. /y/ and /ʏʔ/ are preceded by /tɕ/ and /dʑ/ instead of /t/ and /d/ because of phonotactic restrictions. Since /to/ and /do/ are missing in native Shanghai Chinese phonology, they are replaced by /po/ and /bo/. Correspondingly, the onsets for /ɔ/ and /oʔ/ are also changed to the bilabial stops.

Each syllable was produced twice, ideally represented by two different Chinese characters with different meanings. Altogether, there were 60 stimuli (15 vowels × 2 tone registers × 2 times) as shown in Table 4 that were randomized in the list.

Table 4. Target monosyllabic words

Open syllables		**Glottal-ended syllables**	
T2 (voiceless)	T3 (voiced)	T4 (voiceless)	T5 (voiced)
/tʰɑ/	/dɑ/	/tʌʔ/	/dʌʔ/
太 ‘too’	汰 ‘to wash’	搭 ‘to take’	达 ‘to reach’
泰 ‘Thai’		答 ‘to answer’	踏 ‘to tread’
/pɔ/	/bɔ/	/poʔ/	/boʔ/
宝 ‘treasure’	抱 ‘to hug’	北 ‘north’	薄 ‘thin’
报 ‘to report’	暴 ‘violent’	搏 ‘to fight’	箔 ‘foil’
/po/	/bo/		
把 ‘handlebar’	爬 ‘to climb’		
/tɤ/	/dɤ/	/təʔ/	/dəʔ/
斗 ‘to fight’	头 ‘head’	得 ‘to get’	特 ‘very’
抖 ‘to tremble’	豆 ‘bean’	德 ‘morality’	夺 ‘to rob’
/te/	/de/		
对 ‘right’	队 ‘team’		
/tɛ/	/dɛ/		
胆 ‘courage’	台 ‘table’		
	蛋 ‘egg’		
/tø/	/dø/		
短 ‘short’	团 ‘group’		
	段 ‘segment’		
/ti/	/di/	/tɪʔ/	/dɪʔ/
底 ‘bottom’	电 ‘electricity’	跌 ‘to fall’	碟 ‘dish’
店 ‘store’	地 ‘ground’	滴 ‘drop’	敌 ‘enemy’
/tu/	/du/		
赌 ‘to bet’	度 ‘degree’		
堵 ‘block’	杜 a surname		
/tɕy/	/dʑy/	/tɕʏʔ/	/dʑʏʔ/
举 ‘to raise’	具 ‘tool’	橘 ‘orange’	局 ‘bureau’
句 ‘sentence’	巨 ‘huge’	菊 ‘chrysanthemum’	倔 ‘stubborn’

2. 3 Data collection

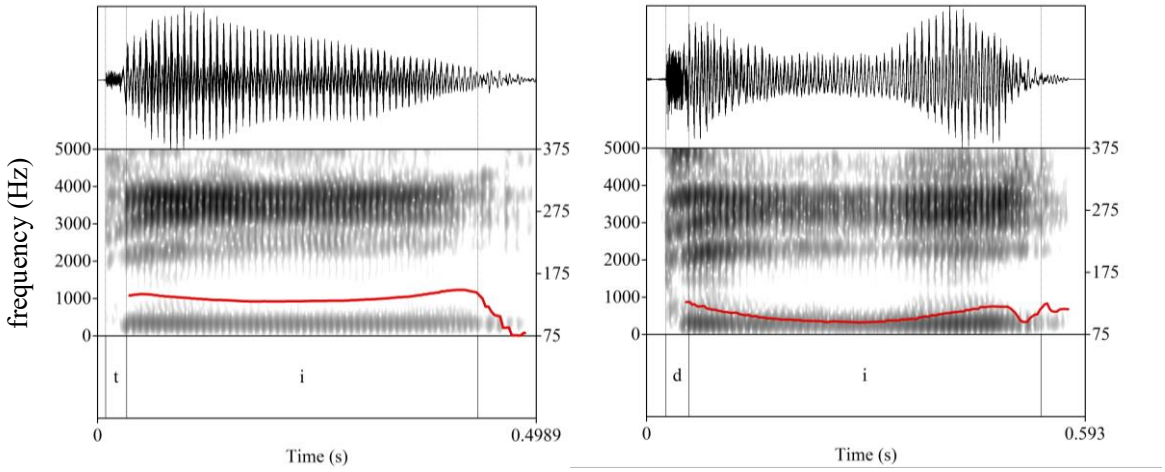
The monosyllabic wordlist reading task was conducted online using the ED (Experiment Designer) software (Vet, 2023), a JAVA-based program. After the consent was obtained, the participants answered four questions about their gender, age, district of living in Shanghai and language background. The participants were required to sit in a quiet room and to wear headsets. Prior to the recording, there was a microphone testing session for the participants to ensure that the recording could be done at a proper volume.

During each trial, the participants first saw a Chinese character in black color in a larger size in the middle of the white screen, and a bi-syllabic word in black color which contained the target Chinese character (in most cases at the initial position of the bi-syllabic word) above

it in a smaller size, as a reference for the correct pronunciation of the target syllable. They were instructed with written Simplified Chinese only to read out aloud that single Chinese character which represents a single mono-syllabic word in Shanghai Chinese rather than the bi-syllabic word, at a normal speed. There was also a blue progress bar at the bottom of the screen showing the progression of the experiment. The target character, bi-syllabic word and the progress bar all disappeared after 1750 ms. The recording started automatically at the same time, with a line of Chinese words in red indicating that it was during the recording session. When the recording finished, the participants pressed the space bar on their keyboard to upload the recording. The uploading took about 500 ms, during which a black fixation cross was shown in the middle of the screen. Then a new trial started.

2. 4 Acoustic measurements

The segmentation and measurement were conducted in Praat (Boersma & Weenink, 2022). Each recording was first clipped automatically by removing the silence according to the intensity with the following setting: 50 Hz minimal pitch, 10 ms time step, -25 dB silence threshold, 100 ms minimal silent interval duration and 75 ms minimal sounding interval duration. To investigate the duration difference between checked and unchecked tones in their citation form, the duration of each vowel was measured manually. The beginning of the vowel is the zero-crossing where the waveform turns from irregular to regular and periodic and starts recurring with a certain shape. The end point is the zero-crossing at which the periodic waveform shape becomes unrecognizable, with the formants fading on the spectrogram. Figure 1 shows the examples of the clipped recording and the extraction of the vowel in the syllable /ti/ (T2, Fig. 1a) and /di/ (T3, Fig. 1b) produced by a male speaker. Further analysis of the phonation type and vowel space was only on the vowel part in each recording.



(1a) (1b)

Fig. 1. The waveform, spectrogram, and fundamental frequency (*f*0) contour of the automatically clipped recordings of syllable /ti/ (T2, Fig. 1a) and /di/ (T3, Fig. 1b) produced by a male native speaker and the illustration of vowel extraction.

In terms of vowel space, the frequencies of the F1 and F2 were obtained using the *Sound: To Formant (burg)...* function in Praat based on linear predictive coding with the following

parameters: 5 ms time step, 5500 Hz as the default formant ceiling for females for the first five formants and 5000 Hz as the default formant ceiling for males for the first five formants, 25 ms window length and 50 Hz pre-emphasis. Since /d/ leads to vowel space reduction at the beginning part of the vowel in Shanghai Chinese (Chen, 2008) and many Germanic languages as English, German and Danish (Stevens & House, 1963; Hillenbrand et al., 2001; Steinlen, 2002), probably as the result of articulatory configuration, we got the F1 and F2 information at the midpoint and upper quartile (i.e., the 75%-point) of the vowel (see Fig. 2). By doing so, the potential difference caused by the different onsets for practice reasons was also largely avoided. Nevertheless, the automatic formant analysis may still go wrong with some vowels in which the two adjacent formants may be merged for being too close. In this case, the prior knowledge of the vowel space in Shanghai Chinese (Svantesson, 1989; Chen, 2008) was referred to, and the formant analysis was repeated with a different setting, i.e., with a larger number of formants to separate the merged formants.

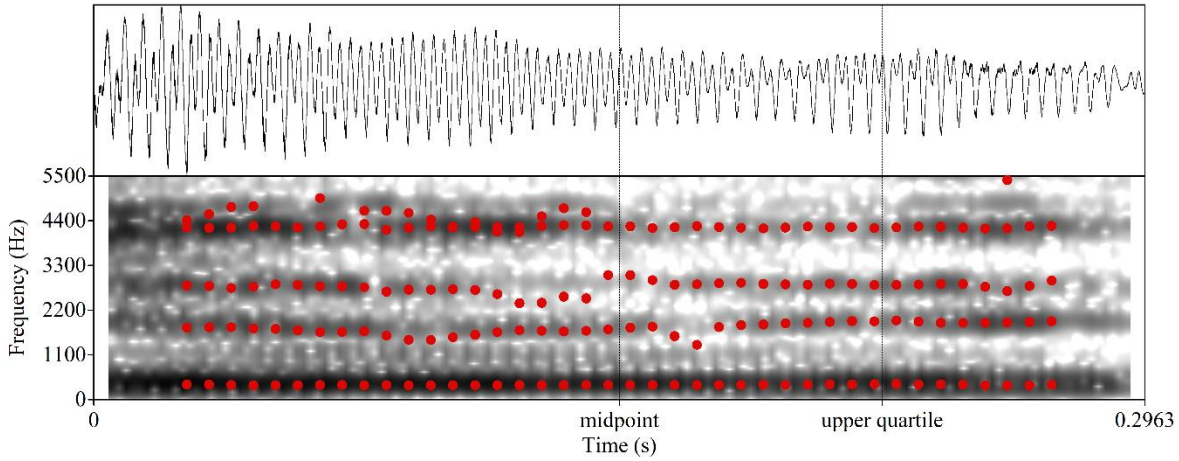


Fig. 2. The waveform and spectrogram of the vowel in syllable /dø/ (T3) produced by a female native speaker. The first two formants were obtained at both the midpoint and the upper quartile.

To examine the phonation type of vowels embedded in low-register tones and high-register tones, there were two measures. First, the H1–H2 difference was measured at the first three complete periods of the vowel to detect the breathiness as according to the data from Gao & Hallé (2017), salient difference between the two tone registers could be found at the beginning and the lower quartile of the vowel, while at the midpoint the difference already became smaller. The harmonics-to-noise ratio (HNR) during stop release, despite being an appropriate and common measure of breathy phonation, was not measured in this study, as (1) the stops are not the focus of this study; and more importantly, (2) H1–H2 difference was reported to be the most sensitive and salient measure of breathiness in Shanghai Chinese (Gao & Hallé, 2017). In practice, a Praat script was used first to detect the pulses in the vowel using *To Pitch (cc)* and *To PointProcess (cc)* functions with the pitch floor set at 75 Hz and the pitch ceiling at 600 Hz, following the default settings. Then it extracted the first three intervals that supposedly correspond to three complete periods, in most cases, ending prior to the lower quartile (i.e., 25%-point) as shown in Fig. 3. In some rare cases of glottal-ended syllables with a checked tone (either T4 or T5) whose duration is too short, each period may last long. Those ending

after the midpoint were discarded as the breathiness may only occur in the first half of the vowel (Cao & Maddieson, 1992; Gao & Hallé, 2017). A spectrum was automatically synthesized based on the sound slice with three complete periods. The power density of the 4th and the 7th bin represents the value of H1 and H2 respectively (as shown in Fig. 4). The script returned a value of –1 for those harmonic values smaller than the values of adjacent bins, indicating non-periodicity in the sound slice. These vowels were then checked manually and decided whether to be adjusted or removed from the voicing analysis case by case, dependent on the duration and position of the non-periodic part.

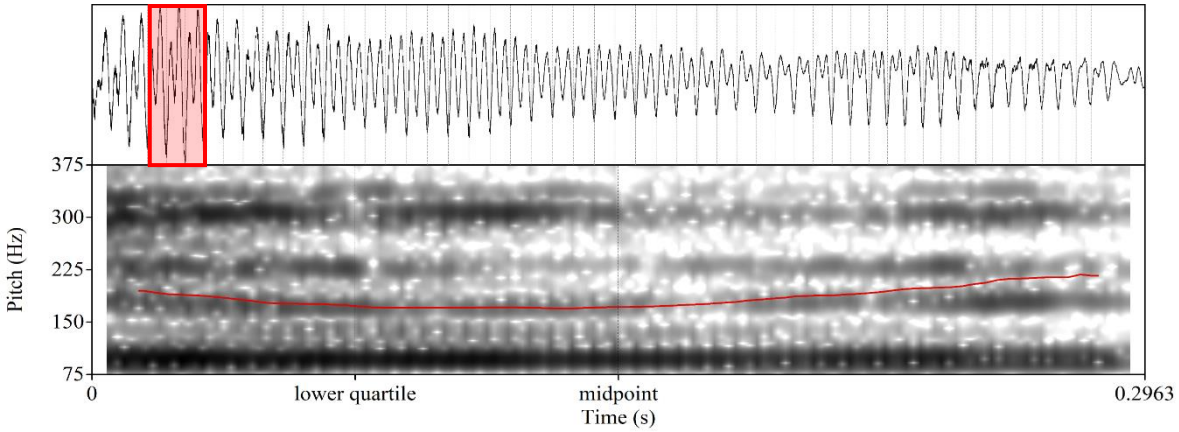


Fig. 3. The waveform and spectrogram of the vowel in syllable /dø/ (T3) produced by a female native speaker (the same sound file as in Fig. 2). The red rectangular shows from which part of the sound the spectrum was extracted.

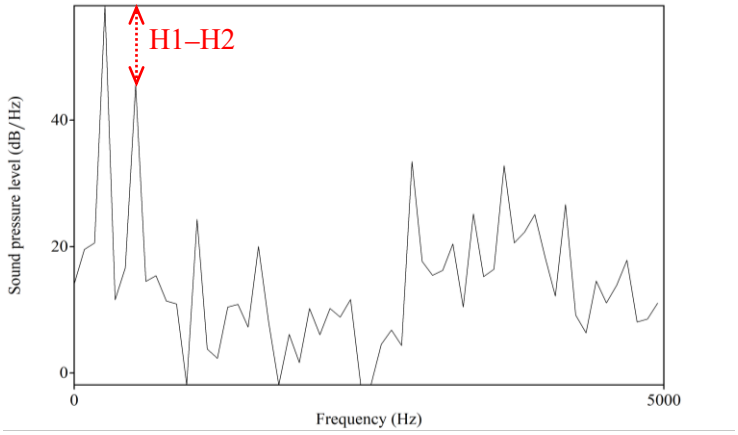


Fig. 4. The spectral slice of three complete periods in the vowel /u/ produced by a female participant that shows the measure of H1–H2 difference.

Second, the creakiness, as strongly related with *f*0, was reflected in an indirect way because of its various ways of realization and positions of occurrence. The number of glottal pulses in the vowel in the point process generated by the Praat script was compared with the number of periods in the point process which was obtained with the following parameters: 0.1 ms shortest period, 20 ms longest period and 1.3 maximum period factor. If the number of intervals is less than the number of glottal pulses (i.e., the number of points) minus one, it means that there were some intervals with extra-low *f*0 (with the frequency less than 75 Hz, the pitch floor) or irregular *f*0 with non-periodic glottal vibration, as in (1):

$$\text{number of intervals} < \text{number of points} - 1 \quad \textbf{(1)}$$

Both *f*0 patterns can be associated with creaky phonation in other Chinese variants (Kuang, 2017; Huang, 2020; Yu & Lam, 2014; Zhang & Kirby, 2020). The difference between the two sides of inequation (1) is henceforth referred to as *f*0 gap, an indicator of the creakiness. All vowels with one or more *f*0 gaps were then manually checked to examine whether a creaky voice occurred in the vowel to rule out those gaps caused by noise or low sound quality, yielding a Boolean variable of creakiness. Fig. 5 shows an exemplar of creaky phonation in T3.

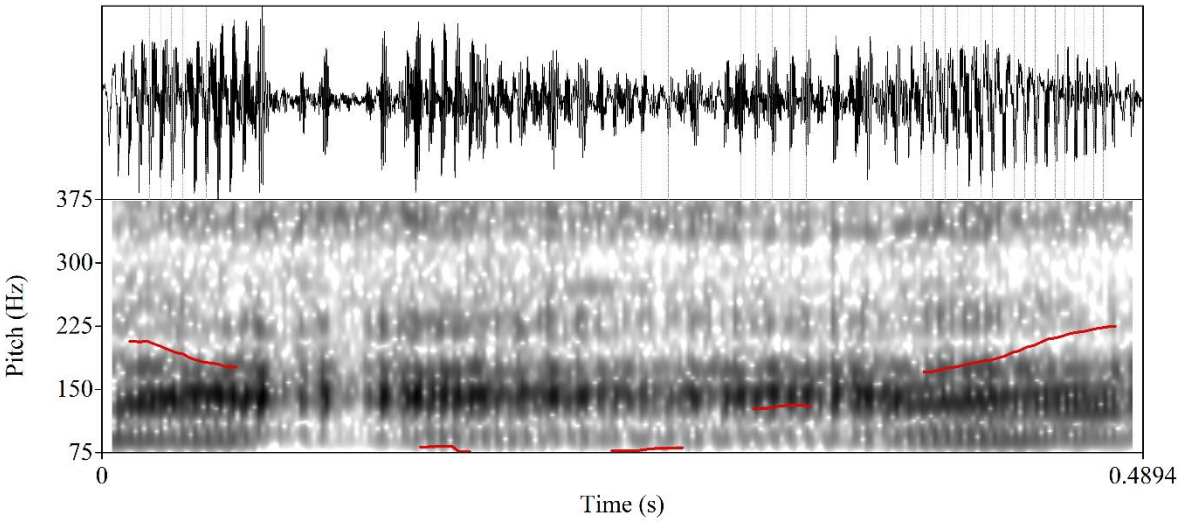


Fig. 5. Vowel /ɑ/ produced with T3 by a female native speaker which demonstrates creaky phonation with 4 visible *f*0 gaps, suggesting that the *f*0 being extra low (less than 75 Hz) or being aperiodic.

2.5 Features, structure and hyperparameter tuning of the LSTM network

An LSTM network consisting of four layers with tuned parameters was built in Python using the *Keras* library (Chollet et al., 2015) running on top of *TensorFlow* framework with an attempt to model the diachronic change of Shanghai Chinese vowels. The model was trained to predict the category into which a vowel most likely falls.

The acoustic measures of duration and vowel space (F1 and F2 at both the midpoint and the upper quartile of each vowel), together with the age and gender of the native-speaking participants functioned as features to be fed into the model to categorize the vowels. Voice quality was not involved in modeling for two reasons: (1) Both breathiness and creakiness are (expected to be) correlated with lexical tones rather than to each vowel, as breathiness derives from the phonemically voiced onsets and creakiness is conditioned by the low *f*0 contour; and (2) There were several values of H1–H2 difference unavailable, which would reduce the size of dataset.

The first (input) layer was a Bidirectional LSTM layer with 256 nodes that employed both dropout (rate = 0.3) and recurrent dropout (rate = 0.3) techniques for regularization. The dropout rates were tuned to mitigate overfitting issues, which might also correspond to the effect of extrinsic factors like language contact on vowel change in the sense that a portion of the input units and activations was randomly dropped out, introducing noise and uncertainty to

the model. By processing the input in both forward and backward directions, this layer helped to better grasp the gradual and nuanced change over time. The input layer was followed by a hidden LSTM layer with 192 nodes, which still returned sequences to retain temporal dependencies and had a dropout rate of 0.1. The third layer was another hidden LSTM layer with 224 nodes. A Dense layer based on a Rectified Linear Unit (ReLU) activation function with 15 nodes representing the 15 vowels in Shanghai Chinese functioned as the output layer.

The Adam optimizer, together with mean squared error as the loss function and accuracy as the metric was used to compile the model. Hyperparameter tuning was carried out by running the *RandomSearch* algorithm for 20 times with 2 executions per trial on the number of nodes (i.e., units) at the first three layers, the number of hidden LSTM layers, the activation function and (recurrent) dropout rates. The tuning phase was based on the training set of 1408 vowels from which 15% of the data were randomly selected as the validation set, a batch size of 256 and 25 epochs. The best parameters as reported above were selected to compile and train the final LSTM network utilizing the entire training set with 50 epochs. With this architecture, it was supposed to effectively capture and predict the dynamics of vowel change patterns.

3. Results

3.1 Overview of the data of the younger generation

146 recordings were discarded for mispronunciation (i.e., a wrong vowel or a wrong lexical tone, judged by a native speaker) or for poor quality of the sound file, composing 10.58% of the 1380 recordings (23 participants × 60 items) produced by the native speakers from the younger generation. The most common mispronunciation was of the monosyllable word 夺 (/dəʔ/, 'to rob') as 20 out of the 23 younger participants produced it as /doʔ/ or /duoʔ/, influenced by its pronunciation in Mandarin Chinese /duo/ with a falling-rising tone. Statistical analysis was conducted on the remaining 1234 recordings.

3.1.1 Vowel duration

The current study aims to examine whether lexical tones make a difference on the duration of vowels in Shanghai Chinese. The four lexical tones involved in the study can be characterized within the two abovementioned dimensions: (1) syllable structure (unchecked tones T2 and T3, coded as +0.5 vs. checked tones T4 and T5, coded as −0.5) and (2) tone register (high tones T2 and T4, coded as −0.5 vs. low tones T3 and T5, coded as +0.5). As shown by Chen (2008), the unchecked tones are on average 1.5 times longer than the checked tones in utterances. This study further investigates whether this pattern holds for words produced in isolation and whether the tone register would also make a difference in vowel duration. Considering gender (with female coded as −0.5 while male as +0.5) and these two predictors, with the random intercepts and slopes (but no random interaction as it would cause the singular fit) of these two predictors per participant, a Linear Mixed-Effects Regression (*lmer*) was built in R 4.3.0 (R Core Team, 2023) with the *lmerTest* package (Kuznetsova et al., 2017) as in (2):

$$lmer\ (Duration \sim gender * syllable\_structure * tone\_register + (syllable\_structure + tone\_register \mid participant)) \quad \textbf{(2)}$$

The main effects of both syllable structure and tone register got confirmed in this model. Unchecked tones (T2 and T3) were on average 202.64 ms longer than checked tones (T4 and T5) when produced as monosyllabic words in isolation, with the 95% confidence interval running from 181.47 to 223.82 ms ($t(21.15) = 18.76$, $p = 1.16 \times 10^{-14}$). Low tones (T3 and T5) were on average 20.92 ms longer than high tones (T2 and T4), with the 95% confidence interval running from 9.96 to 31.8892 ms ($t(22.40) = 3.74$, $p = .00111$). Moreover, their negative interaction was also found with an estimate of −36.81 ms, indicating the effect of syllable structure and that of tone register would partially cancel each other out in the sense that the durational difference between unchecked tones and checked tones were on average 36.81 ms (95% confidence interval running from 25.95 to 47.67 ms) shorter when the tone register was low rather than high ($t(1167.72) = -6.77$, $p = 2.11 \times 10^{-11}$). In other words, the durational difference between T2 and T4 was larger than that between T3 and T5, and that between T5 and T4 was also larger than that between T3 and T2. Post-hoc comparisons performed in the *emmeans* package version 1.8.6 (Lenth et al. 2023) in R using Bonferroni method for multiple pairwise comparisons further showed that T4 was significantly shorter than T5 (estimate = 39.33 ms, $t(44.0) = 5.92$, $p = 4.39 \times 10^{-7}$) but no significant difference between T3 and T2 with an estimated marginal mean of 2.52 ms and the 95% confidence interval between −8.81 to 13.85 ms ($t(25.3) = 0.436$). Therefore, the duration of the four tones under study can be ranked as follows: T3 ≈ T2 > T5 > T4. Fig. 6 and Table 5 illustrate the duration of each vowel. Since the main effect of gender was not detected in the study, duration was reported regardless of the participants' gender. The durational difference among vowels sharing the same syllable structure was rather small, especially within those ending with a glottal coda associated with the checked tones. Open-ended vowels were around 2.4 times as long as their glottal-ended counterparts.

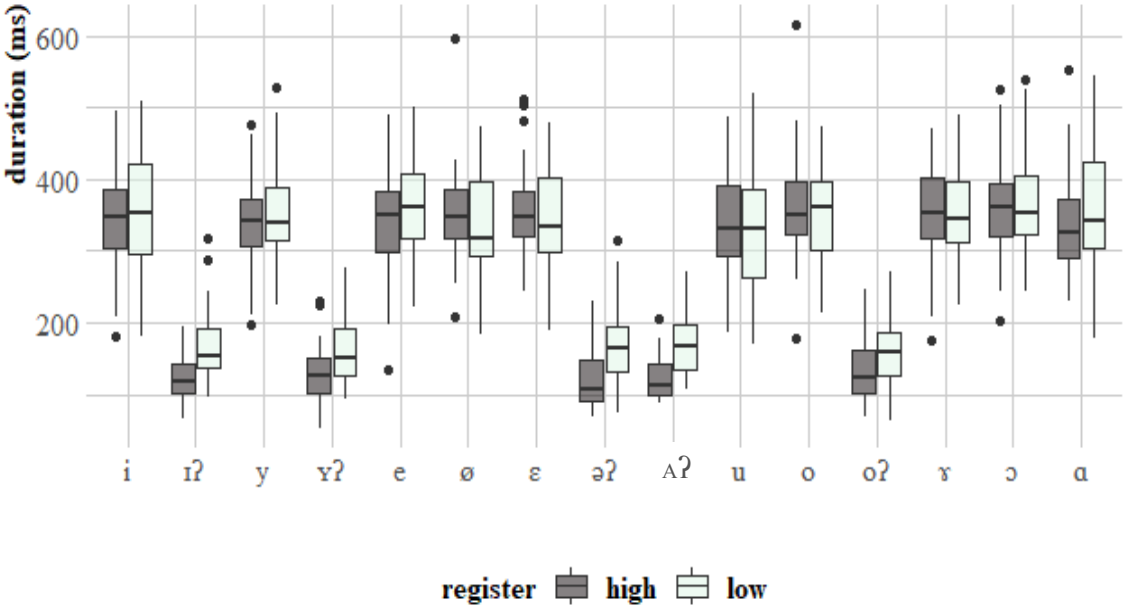


Fig. 6. Boxplot demonstrating the distribution of duration (ms) of the 15 vowels in the two tone registers in Shanghai Chinese produced by the younger generation native speakers ($n = 23$).

Table 5. The mean (and standard deviation) of duration (ms) of each vowel

	front					middle		back				
open	i	y	e	ø	ɛ			u	o	ɤ	ɔ	ɑ
ended	353	344	354	346	349			332	355	354	364	348
	(78)	(63)	(72)	(64)	(68)			(74)	(69)	(65)	(68)	(75)
glottal	ɪʔ	ʏʔ				əʔ	ʌʔ		oʔ			
ended	146	145				142	144		146			
	(45)	(45)				(54)	(45)		(48)			

3.1.2 Vowel space

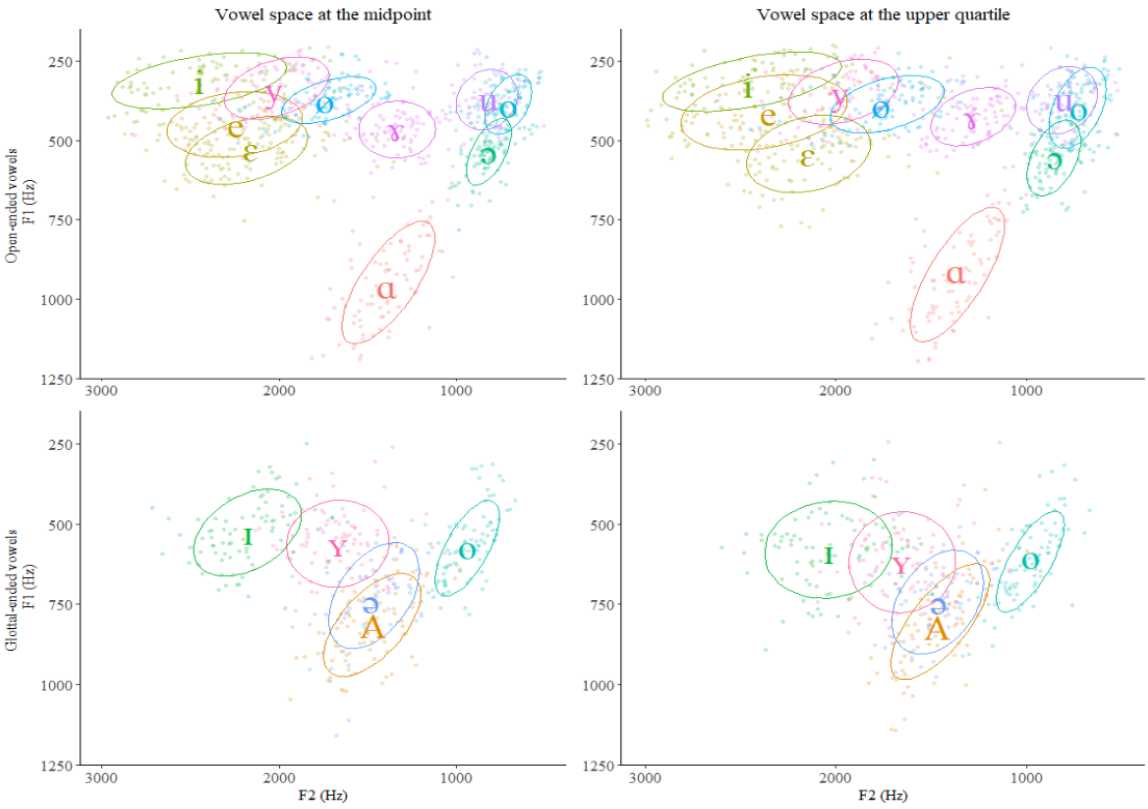


Fig. 7. Scatterplot of the space of the 15 vowels in Shanghai Chinese produced by the younger generation native speakers (n = 23). The two figures on the left column illustrate the vowel space obtained at the midpoint, and the two on the right illustrate that at the upper quartile. The border of each vowel was marked with 1-σ concentration ellipses.

The F1 and F2 values were measured at the midpoint and the upper quartile of each vowel. Fig. 7 visualizes the vowel space at the two timepoints with the mixed data from female and male native speakers. It shows clearly that the vowel space of glottal-ended vowels was more centralized than that of open-ended vowels. More importantly, the distribution of these two sets of vowels was complimentary, as open-ended vowels were in general more peripheral, leaving the central area empty, while glottal-closed vowels mainly occupied that area. As for open-ended vowels, the non-high and non-low vowels were elevated in general, approaching the height of high vowels /i, y, u/. As a result, the vowel space of front vowels /y, e, ɛ, ø/ was overlapping. Back vowels /u/ and /o/ were highly overlapping, especially at the upper quartile.

In terms of glottal-ended vowels, the space of /əʔ/ and /ʌʔ/ was also highly overlapping.

To investigate whether the vowel space changed significantly over time in production, paired samples *t*-tests (15 for F1 values and 15 for F2) with aggregated data per participant were run for each vowel in R 4.3.0 (R Core Team, 2023). The mean F1 and F2 of each vowel produced by female and male speakers at the two timepoints and the result of the paired *t*-tests were summarized in Table 6 and 7. The series of paired *t*-tests showed that the height of /i, ø, ɛ, ɔ, ɪʔ, ʏʔ, oʔ/ reduced whilst that of /e, ɤ, ɑ/ rose in the second half of the vowel. In terms of frontness, however, only /ɤ, y, ɪʔ/ exhibited a tendency of moving backwards whilst /ɔ, oʔ, ø, e/ moved forwards in the second half of the vowel. Among these notable movements, only the vertical movements of /ɪʔ, ʏʔ, e, ɤ/ and the horizontal movements of /y, ɪʔ/ were significant when compared to the conservative Bonferroni-corrected alpha level (the original alpha level of .05 divided by 15, the total number of tests, performed to control the family-wise error rate).

Table 6. The mean F1 (Hz) of each vowel and the result of paired *t*-tests

	midpoint		upper quartile		paired *t*-test result		
	Female	Male	Female	Male	*t*-value	*p*-value	95% CI
i	354	273	366	284	$t(21) = -2.35$	.02882*	(−20.98, −1.27)
y	378	300	395	307	$t(22) = -1.22$	.235	(−24.30, 6.29)
u	394	343	402	331	$t(21) = 0.38$	.7062	(−16.05, 23.28)
ø	411	336	432	343	$t(22) = -2.69$	.01327*	(−24.51, −3.19)
o	415	358	446	347	$t(21) = -0.95$	.3509	(−26.87, 9.97)
e	482	418	458	365	$t(22) = 4.35$	.000257**	(19.80, 55.89)
ɤ	487	433	460	384	$t(22) = 3.36$	.002813**	(14.38, 60.68)
ɛ	557	485	573	507	$t(22) = -2.65$	.01469*	(−32.81, −3.99)
ɔ	567	496	580	521	$t(22) = -2.83$	.009699*	(−30.99, −4.79)
ɪʔ	583	464	626	537	$t(22) = -6.10$	3.876×10^{-6}**	(−76.97, −37.91)
ʏʔ	611	512	666	560	$t(22) = -5.17$	3.457×10^{-5}**	(−72.68, −31.09)
oʔ	626	514	668	537	$t(22) = -3.26$	.003556*	(−60.11, −13.40)
əʔ	773	706	770	715	$t(22) = -0.45$	.6545	(−26.16, 16.77)
ʌʔ	841	787	860	780	$t(22) = -0.49$	.6302	(−49.32, 30.52)
ɑ	1012	901	959	859	$t(22) = 2.87$	.008946*	(13.67, 85.09)

Note. The vowels are ranked following the ascending order of values in the second column. Negative *t* values indicate that the vowel height descended from the midpoint to the upper quartile. In the second last column, one asterisk means the *p*-value is small than .05 and two asterisks mean the *p*-value is small than the Bonferroni-adjusted significance level, i.e., < .05/15.

To further examine the factors affecting vowel space, four *lmer* models (3-6) were built, whose outcome (dependent variable) was F1 at the midpoint, F1 at the upper quartile, F2 at the midpoint, and F2 at the upper quartile respectively. The two predictors related to lexical tones (i.e., syllable structure being open or closed and tone register being high or low) and gender of participants (female coded as +0.5 while male coded as −0.5) were included as fixed effects in all four models. Vowel highness was included in the two models on F1 as an ordinal variable

with three levels: High, Mid and Low even though it is highly correlated with F1 values. Backness was also included in the two models on F2 with three levels: Front, Mid and Back despite its correlation with F2 values. These two predictors were included because we were interested in how they would interact with other main effects. Table 8 concludes the significant vowel space movement (indicated by the arrows) from the midpoint to the upper quartile detected in the paired *t*-tests and its direction and shows the highness and backness labels that each vowel was assigned. Since the glottal-ended vowels demonstrated a more centralized distribution than the open-ended vowels, making it difficult to standardize the two vowel sets, highness and backness were labelled with reference to the vowel position in the International Phonetic Alphabet (IPA) chart (International Phonetic Association, 2015). As for random intercepts and slopes, both predictors related to lexical tones were included, unless convergence or singular fit issues occurred.

Table 7. The mean F2 (Hz) of each vowel and the result of paired *t*-tests

	midpoint		upper quartile		paired *t*-test result		
	Female	Male	Female	Male	*t*-value	*p*-value	95% CI
o	720	691	742	708	$t(21) = -1.31$	.2038	(−48.26, 10.93)
u	834	796	811	787	$t(21) = 0.87$	.3962	(−23.01, 55.86)
ɔ	853	775	886	806	$t(22) = -2.34$	.02883*	(−53.85, −3.23)
oʔ	1008	862	1054	899	$t(22) = -2.84$	.009451*	(−67.04, −10.49)
ɤ	1338	1337	1300	1260	$t(22) = 2.52$	.01932*	(9.93, 101.38)
ɑ	1475	1305	1441	1287	$t(22) = 1.68$	.1064	(−5.75, 55.36)
ʌʔ	1524	1409	1523	1402	$t(22) = -0.18$	.8595	(−28.31, 23.81)
əʔ	1572	1383	1550	1375	$t(22) = 1.03$	.3128	(−13.98, 41.72)
ʏʔ	1734	1592	1712	1586	$t(22) = 1.07$	.2973	(−13.83, 43.16)
ø	1856	1625	1891	1633	$t(22) = -2.11$	.04674*	(−48.81, −0.39)
y	2174	1893	2122	1832	$t(22) = 4.66$	.0001208**	(31.18, 81.21)
ɛ	2274	2030	2251	2016	$t(22) = 0.68$	.5055	(−28.55, 56.21)
ɪʔ	2303	2027	2119	1936	$t(22) = 5.01$	5.14×10^{-5}**	(82.79, 199.73)
e	2344	2132	2477	2212	$t(22) = -2.17$	.04123*	(−209.08, −4.65)
i	2575	2321	2579	2330	$t(21) = -0.41$	.6827	(−51.88, 34.85)

Note. The vowels are ranked following the ascending order of values in the second column. Negative *t*-values indicate that the vowel moved towards the front from the midpoint to the upper quartile. In the second last column, one asterisk means the *p*-value is small than .05 and two asterisks mean the *p*-value is small than the Bonferroni-adjusted significance level, i.e., < .05/15.

$$lmer\ (F1\_midpoint \sim gender * syllable\_structure * tone\_register * vowel\_highness + (syllable\_structure + tone\_register \mid participant)) \quad \textbf{(3)}$$

The model of F1 values at the midpoint of each vowel revealed the main effects of gender, syllable structure and vowel highness. The estimate difference of F1 between female and male native speaking participants was 76.55 Hz (with the 95% confidence interval running from 40.62 to 112.48 Hz), indicating that the vowels produced by female native speakers were

significantly lower than those produced by male native speakers ($t(22.87) = 4.18$, $p = .000367$). The vowel height of glottal-ended vowels was significantly lower than open-ended vowels, with an estimated difference of −109.937 Hz (the 95% confidence interval from −134.26 to −85.61 Hz), which means that glottal-ended vowels in general have a lower vowel height than open-ended vowels ($t(24.89) = -8.86$, $p = 3.63 \times 10^{-9}$). Both contrasts of highness turned out to be significant, with a p-value for mid – high and low vowel contrast of 2.50×10^{-19} and that for high – low vowel contrast smaller than 10^{-308}. Middle vowels /e, ø, ɛ, ɤ, ɔ, əʔ/ exhibited F1 values much closer to high vowels than to low vowels, with $t(1155.82) = 9.15$. The model also revealed two two-way interactions, both were between vowel height and syllable structure, with $t(1153.36) = 18.35$ and $p = 3.69 \times 10^{-66}$ for the height contrast between mid and high/low vowels and with $t(1152.38) = 22.49$ and $p = 3.84 \times 10^{-93}$ for the height contrast between high and low vowels.

Table 8. Space movement, its direction, and the highness and backness labels of each vowel

front					middle		back				
i	y	e	ø	ɛ			u	o	ɤ	ɔ	ɑ
high↓	high	mid↑	mid↓	mid↓			high	high	mid↑	mid↓	low↑
front	front→	front←	front←	front			back	back	mid→	back←	mid
ɪʔ	ʏʔ				əʔ	ʌʔ		oʔ			
high↓	high↓				mid	low		high↓			
front→	front				mid	mid		back←			

Note. Highness was coded as follows in the order of High, Low and Mid: (1/3, 1/3, −2/3) for the first contrast and (−0.5, 0.5, 0) for the second contrast. Backness was coded as follows in the order of Back, Front and Mid: (1/3, 1/3, −2/3) for the first contrast and (−0.5, 0.5, 0) for the second contrast.

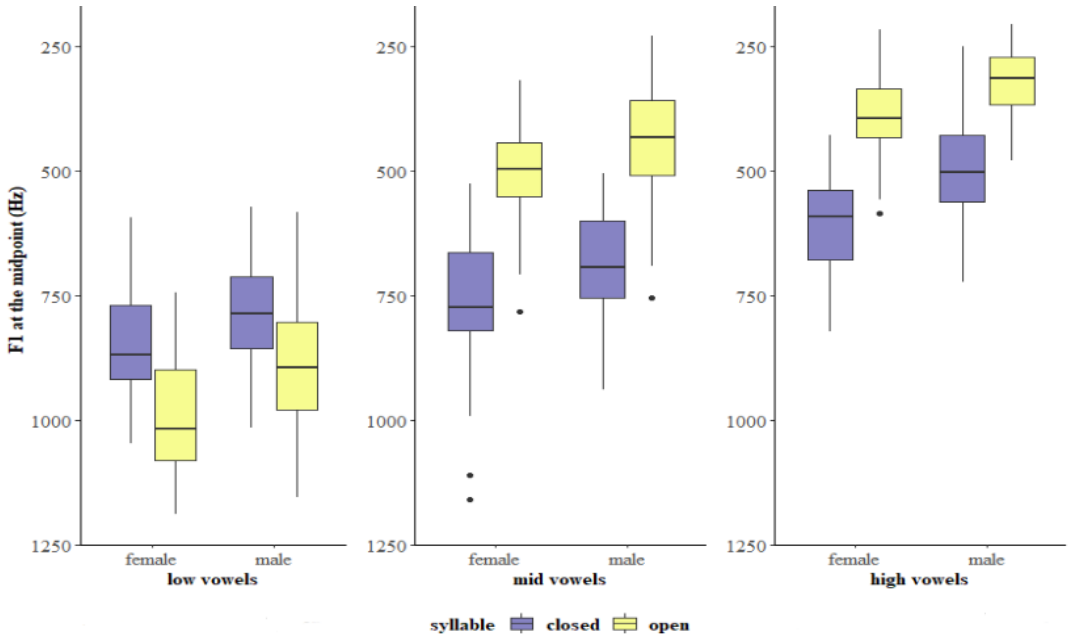


Fig. 8. Boxplot showing the three-way interaction among gender, syllable structure and vowel highness at the midpoint of vowels.

Three three-way interactions were found in this model. The first was a crossover interaction among gender, syllable structure and the highness of low and high vowels (t(1152.38) = 3.54 and p = .000418), suggesting the significant difference in F1 between female and male participants held across syllable structure and the highness of vowels (Fig. 8), while the effect of highness and that of syllable structure were opposite. Most interestingly, the other two three-way interactions involved tone register apart from syllable structure and vowel highness. Post-hoc comparisons further revealed that the open-ended low vowel, i.e., /ɑ/ had a significantly higher F1 value when the tone register was low rather than high. In other words, /ɑ/ produced in T3 has a lower height than /ɑ/ produced in T2. The estimated difference was 43.23 Hz, and the 95% confidence interval ranged from 6.40 to 80.44 Hz with t(800.3) = 2.30 and p = .0218.

$$lmer\ (F1\_0.75point \sim gender * syllable\_structure * tone\_register * vowel\_highness + (syllable\_structure \mid participant)) \quad \textbf{(4)}$$

At the upper quartile of the vowel duration, one more main effect was observed. The vowels with low-register tones (i.e., T3 and T5) in general had a higher F1 value than those with high-register tones (i.e., T2 and T4), with an estimated difference of 17.45 Hz lying in the middle of the 95% confidence interval that ran from 2.92 to 31.99 Hz (t(1171.67) = 2.35, p = .01874). It also revealed the negative interaction between gender and register. The female – male difference in F1 was on average 38.52 Hz (95% confidence interval running from 9.45 to 67.58 Hz, t(1171.67) = –2.60, p = .00952) smaller when the vowel was produced with a low tone rather than a high tone.

In terms of the three-way interactions, however, the model with F1 data at the upper quartile showed different results in the sense that it did not reveal gender × syllable structure × highness interaction. Instead, it showed the significant interaction of gender × tone register × highness. Post-hoc comparisons found the significant female – male difference for high vowels produced in high tones (t(30.1) = 3.85, p = .0006) and low tones (t(31.1) = 3.95, p = .0004), for low vowels produced in high tones (t(70.2) = 4.23, p = .0001) and for mid vowels (both non-high and non-low vowels) produced in high tones (t(44.6) = 2.679, p = .0103). They also revealed another significant difference in F1 between high and low vowels for glottal-ended high vowels /ɪʔ, ʏʔ, oʔ/ (t(1171) = –2.96, p = .0031) on top of that for open-ended low vowel /ɑ/, which was already shown at the midpoint.

$$lmer\ (F2\_midpoint \sim gender * syllable\_structure * tone\_register * vowel\_backness + (syllable\_structure \mid participant)) \quad \textbf{(5)}$$

The two models on F2 values showed less complicated information. The model in which F2 at the midpoint of a vowel revealed the main effects of gender and backness between front and back vowels. The F2 value of vowels produced by female native speakers was on average 153.40 Hz higher than that produced by male native speakers (with the 95% confidence interval from 80.26 to 226.52 Hz, t(21.81) = 4.11, p = .000467. It means that the female speakers in general produce the vowels fronter than male speakers. The front vowels were much fronter

than the back vowels with an estimated difference of 1153.14 Hz lying in the middle of the 95% confidence interval running from 1114.68 to 1191.60 Hz (t(1177.38)= 58.76, $p < 10^{-308}$). Unlike the middle vowels in terms of F1 (i.e., highness) whose height was close to high vowels, the middle vowels in backness did not differ significantly from the average F2 of back and front vowels.

Apart from these two main effects, the model also revealed the interaction between gender and backness as well as between syllable structure and backness. The F2 difference between back and front vowels was on average 130.70 Hz larger for female native speakers than for male native speakers (with the 95% confidence interval between 53.77 and 207.62 Hz, t(1177.38) = 3.33, p = .000895). This difference was also greater for open-ended vowels than for glottal-closed vowels with an estimated difference of 369.00 Hz (95% confidence interval between 292.08 and 445.93 Hz, t(1176.66) = 9.40, $p = 2.69 \times 10^{-20}$). The F2 difference between front/back vowels and mid vowels was also on average 136.03 Hz (95% confidence interval from 68.98 to 203.08 Hz) larger for open-ended vowels than for glottal-closed vowels, with t(1172.47) = 3.98 and $p = 7.42 \times 10^{-5}$.

$$lmer\ (F2\_0.75point \sim gender * syllable\_structure * tone\_register * vowel\_backness + (syllable\_structure \mid participant)) \quad \textbf{(6)}$$

The last model was on F2 at the upper quartile point. Compared with the results from the model based on formular (5) on F2 at the midpoint, the model based on formula (6) showed an extra main effect of backness between front/back vowels and mid vowels as well as an extra three-way interaction of gender × syllable structure × backness (front vs. back vowels). The main effect meant that middle vowels were on average 43.27 Hz (95% confidence interval from 9.45 to 77.09 Hz) closer to front vowels than to back vowels at the upper quartile, with t(1171.87) = 2.51, p = .01128. Its positive interaction (t(1176.00) = 2.53, p = .01147) with gender and syllable structure was further examined using post-hoc comparisons, which showed that the female – male difference widely held across the following four situations: (1) a glottal-ended back vowel, i.e., /oʔ/, with t(83.2) = 2.20, p = .0303; (2) glottal-ended front vowels, i.e., /ɪʔ, ʏʔ/, with t(39.0) = 2.81, p = .0078; (3) glottal-ended mid vowels, i.e., /əʔ, ɐʔ/, with t(44.6) = 2.37, p = .0223; (4) open-ended front vowels, i.e., /i, y, e, ø, ɛ/, with t(30.4) = 6.27, $p = 6.33 \times 10^{-7}$. In brief, the gender difference turned out to be significant in all five glottal-ended vowels but only in front open-ended vowels at the upper quartile.

3.1.3 Phonation type

The two non-modal phonation types: breathy voicing and creaky voicing are the focus of this study. The former was measured with the H1–H2 difference at the first three periods of the vowel, and the latter was indirectly shown by the number of *f*0 gaps, which was manually checked and coded as a binary variable: creakiness. For the breathiness, model (7), an *lmer* model in which H1–H2 difference functioned as outcome was constructed as follows:

$$lmer\ (H1 - H2\_difference \sim gender * syllable\_structure * tone\_register + (syllable\_structure + tone\_register \mid participant)) \quad \textbf{(7)}$$

Model (7) unexpectedly showed the main effect of syllable structure and its interaction with gender but no main effect of tone register. Open-ended vowels demonstrated higher H1–H2 difference than glottal-ended vowels, with an average sound pressure level of 1.55 dB/Hz, falling in the 95% confidence interval between 0.50 and 2.60 dB/Hz ($t(20.19) = 2.90$, $p = .00876$). It had a positive interaction with gender, indicating that the difference between the two harmonics was on average 3.36 dB/Hz larger for female than for male, with the 95% confidence interval running from 1.27 to 5.46 dB/Hz, $t(20.19) = 3.15$ and $p = .00506$. These findings show that glottal-ended vowels were produced with a lower spectral tilt and thus were tenser than open-ended vowels, and this gap was larger for female native speakers than for male native speakers. The effect size of difference between high- and low-register vowels was quite small, with the estimated mean of –0.16 dB/Hz and the 95% confidence interval from –1.84 to 1.52 dB/Hz ($t(23.23) = –0.19$).

$$glmer\ (creakiness \sim gender * syllable\_structure * tone\_register + (tone\_register \mid participant),\ family = "binomial") \quad \textbf{(8)}$$

As creakiness was a binary outcome, a generalized linear mixed-effects regression (*glmer*) model was built according to formula (8). It revealed the main effects of both tone register and syllable structure as well as their interaction. Low-register tones were on average 1.78 times (i.e., the estimated odds = 1.78, while the original estimate was 0.57) more likely to be produced with an extra-low *f*0 or irregular *f*0 than high-register tones. The 95% confidence interval ran from 1.04 to 3.03, $z = 2.10$ and $p = .0356$. Open-ended syllables were on average 1.90 times (i.e., the estimated odds = 1.90, while the original estimate was 0.64) more likely to be produced with a creaky voice than glottal-ended syllables. The 95% confidence interval ran from 1.22 to 2.95, $z = 2.85$ and $p = .0044$. Their positive interaction had the estimated odds of 3.07 and the 95% confidence interval between 1.27 and 7.43 ($z = 2.49$, $p = .0126$), suggesting that low-register open-ended vowels (i.e., vowels with T3) was on average 3.07 times more likely to carry a creaky voice than high-register glottal-ended vowels (i.e., vowels with T4). Post-hoc comparisons only exhibited a significant high – low register difference in open-ended syllables, with the estimated odds = –3.13 (the original estimate = –1.14), $z = –4.27$ and $p = 1.89 \times 10^{-5}$.

In summary, the models did not find the expected significant H1–H2 difference between high and low tone vowels measured at the first three detectable periods in the younger generation. However, we found out that vowels assigned with a low tone were more likely to be produced with an extra-low *f*0 or irregular *f*0 than those assigned with a high tone. It was unexpected that syllable structure played a noticeable role in both measures of voice quality.

3.2 LSTM modeling of the diachronic change over generation

Altogether 1565 recordings served as the input to the LSTM network in which 331 recordings were from the six elder generation native speakers aged between 40 and 60 at the time when the recording took place (excluding 29 recordings discarded due to poor quality of the sound file or to mispronunciation, judged in the same standard as the recordings from the younger generation). 157 (10%) were randomly selected to compose a test set, while the remaining 1408

functioned as the training set. The evaluation of this model was based on the test set with precision, recall and F-score, which are common measures in machine learning.

The LSTM network had an overall precision of 0.6026, which is the ratio of true positives (the number of correctly predicted vowel categories) to the sum of true positives and false positives (i.e., Type I error). It meant that vowels categorized as the target vowel had a 60.26% chance to be categorized correctly. In other words, there was a 39.74% chance that vowels predicted to be in the same category were actually from other categories. The model had an overall recall of 0.5987, the ratio of true positives to the sum of true positives and false negatives (i.e., Type II error), suggesting that the model had a chance of 40.13% to miss a vowel that should be categorized as the target vowel. The quite close values of precision and recall suggest that the model showed a well-balanced performance with the two types of errors. F-score, the harmonic mean of the precision and recall, was 0.6006 in this case.

Diving deeper into the performance of the LSTM model on the test data, we checked the distribution of the vowel categorizes assigned by the model and their actual categories, as summarized in Table 9. The two vowel sets (open-ended vowels vs. glottal-ended vowels) were usually not confused, probably because of the duration feature. The exceptions found in the test set were that /ɪʔ/ was mistakenly predicted to be /ɛ, ø/, /ʏʔ/ was predicted to be /y, ø, ɤ/ and /oʔ/ was predicted to be /ɔ/ with relatively close F1 and F2 values. Among glottal-ended vowels, /ɪʔ/ was often mistaken as /ʏʔ/, /ʏʔ/ as /oʔ/, /əʔ/ as /ɐʔ, oʔ/ while /ɐʔ/ as /oʔ/, which could also be explained by their overlapping vowel space. Open-ended /i, ɛ, ø, ɑ, ɤ, ɔ, u/ were labeled correctly in most cases in our limited test dataset. However, the model did not perform well with vowels /y, e, ɯ/, in the sense that more than half of /y/ was predicted by the model as /ø/, many in the /e/ class were labelled as /i/, and /ɯ/ was mostly predicted as /u/.

Table 9. Confusion matrix of the vowel categories predicted by the LSTM network

		predicted category														
		ɪʔ	ʏʔ	əʔ	ɐʔ	oʔ	i	y	e	ɛ	ø	ɑ	ɤ	ɯ	ɔ	u
actual vowel	ɪʔ	**3**	6	0	0	0	0	0	0	1	1	0	0	0	0	0
	ʏʔ	0	**4**	1	0	3	0	1	0	0	1	0	1	0	0	0
	əʔ	0	1	**1**	6	3	0	0	0	0	0	0	0	0	0	0
	ɐʔ	0	1	0	**4**	2	0	0	0	0	0	0	0	0	0	0
	oʔ	0	0	0	0	**8**	0	0	0	0	0	0	0	0	1	0
	i	0	0	0	0	0	**13**	1	0	0	0	0	0	0	0	0
	y	0	0	0	0	0	0	**3**	0	0	6	0	0	0	0	0
	e	0	0	0	0	0	6	2	**5**	1	1	0	0	0	0	0
	ɛ	0	0	0	0	0	0	0	3	**7**	2	0	0	0	0	0
	ø	0	0	0	0	0	1	0	0	0	**9**	0	0	0	0	0
	ɑ	0	0	0	0	0	0	0	0	0	0	**5**	0	0	0	0
	ɤ	0	0	0	0	0	0	0	0	0	0	0	**9**	1	0	0
	ɯ	0	0	0	0	0	0	0	0	0	0	0	0	**3**	0	8
	ɔ	0	0	0	0	0	0	0	0	0	0	0	0	0	**9**	1
	u	0	0	0	0	0	0	0	0	0	0	0	0	0	1	**11**

Because the LSTM model obtained a mediocre F-score of 0.6006 on a small test set, we do not proceed with predicting Shanghai Chinese vowel shift tendency in the coming years in the current study. Nevertheless, the patterns revealed from the performance and error analysis of this LSTM network are discussed in detail in the next section.

4. General discussion

The present study investigated the acoustic realization of vowels in Shanghai Chinese in the youngest generation born after 1995 and to tackle their diachronic change over decades with the data obtained from a monosyllabic word-reading task. The statistical results showed that the duration of open-ended vowels is on average around 2.4 times the duration of glottal-ended vowels when produced in isolation and that both predictors related to lexical tones, i.e., syllable structure and tone register affect vowel duration. The duration of the four lexical tones in the same vowel can be ranked as: T3 ≈ T2 > T5 > T4 (the 95% CI for the durational difference between T3 and T2 is from −8.81 to 13.85 ms). In terms of vowel space, the fact that glottal-ended vowels are distributed more to the center whilst the open-ended vowels are more peripheral got confirmed with the consistent interaction between syllable structure and vowel highness or backness across all four *lmer* models on vowel space. The height of non-high and non-low vowels is closer to the high vowels rather than the low vowels. With regard to phonation type, the younger generation seems not to produce low-register vowels with more airflow coming out, at least not systematically different from high-register vowels that are supposed to be produced with a modal voice. On the other hand, we found out that low-register vowels are often accompanied by extra-low or irregular *f*0 value, which often functions as the sign of a creaky voice. Meanwhile, syllable structure plays a role in both measures of voice quality.

4.1 Vowel duration

As mentioned above, open-ended vowels are on average 2.4 times as long as glottal-ended vowels when produced in their citation form in isolation. This durational difference is greater than that 1.5-time difference reported by Chen (2008) with monosyllabic words embedded in carrier sentences, confirming our hypothesis in Section 1.3. The discrepancy in the two studies is probably caused by the prosodic prominence (Chen, 2008; Crystal & House, 1990) as the monosyllabic words produced in isolation are in default under focus or stressed and by the fewer segments in a prosodic unit (Crystal & House, 1990).

Under the same condition of being produced in isolation, the vowel duration is strongly affected by lexical tones in Shanghai Chinese, not only the factor of syllable structure but also the factor of tone registers. The effect of syllable structure (checked tone with a glottal stop ending vs. unchecked tones with a sonorant ending in our case) has also been found in several other languages like Cantonese (Wong & Chan, 2018), Min Chinese (Chai & Ye, 2022) and Burmese (Gruber, 2011) although the realization of the syllable-end coda varies among languages. In some languages, the checked tones may be accompanied with phonation cues. In White (H)mong, for example, the checked tone has a creaky phonation that is decisive to duration (Esposito, 2012; Garellek & Esposito, 2021). The durational difference between open-

ended and glottal-ended vowels, however, probably strengthens rather than dominates the phonological distinction between checked and unchecked tones. After all, it is the phonation type instead of duration that serves as the primary cue in perception of the lexical tones with similar *f*0 contour patterns in many tonal languages, as the case in Vietnamese (Brunelle, 2009), Cantonese (Yu & Lam, 2014) and Green (H)mong (Andruski, 2006).

This study also found an important role played by tone register in determining Shanghai Chinese vowel duration, which is a new finding not mentioned in previous literature. Vowels with low tones are significantly longer than vowels with high tones. It is also a universal pattern which has been observed in other tonal languages like Mandarin Chinese (Wu & Kenstowicz, 2015) and Southern Min Chinese (Zee, 1978). One possible explanation would be that it is easier and faster for the pitch to lower than to rise (Zee, 1978; Ohala & Ewan, 1973), accounting for the fact that T5 (a short low-rising tone) is much longer than T4 (a short high-falling tone). T2 and T3 share a very similar pitch contour pattern, hence no significant durational difference.

4.2 Vowel space

Associated with a checked tone (either T4 or T5), glottal-ended vowels display a more centralized vowel space. It is plausible to say that the short duration of glottal-ended vowels may be the cause of vowel space centering from the articulatory side, as to reach a more centralized position in the oral cavity after full closure is easier and takes less time than to reach a more peripheral position. Such a contrast in both duration and vowel space also exists in other languages, as the /i/ – /ɪ/ contrast in English which started as a pure length contrast between /iː/ – /i/ in Middle English but later turned into a more/less central vowel space distinction. This historical change provides evidence that it is more likely that vowel shortness leads to vowel centralization rather than the other way around (Delattre, 1962; House,1961).

Vowel space also changed mildly from the midpoint to the 75% point, as evidenced by the series of paired samples *t*-tests, yet no universal pattern was found except that there seems to be less significant movement horizontally (in backness) than vertically (in highness). Nevertheless, difference was observed in /ɤ, ɔ, ɪʔ, oʔ/ and the newly re-emerged phoneme /e/ in both directions. /ɤ/ moves from a more central position closer to the center of the vowel space, where the glottal-ended schwa locates to a higher and backer position closer to /u/. This pattern is unlikely to be the result of vowel space coordination or compromise for maximal distinction between phonemes as it is the only open-ended vowel occupying the central area. On the contrary, in both cases that /ɔ/ moved downward and forward and /e/ moved upward and forward, they are surrounded by other vowels. The former is close to /u, o/ while the latter partially overlaps with /ɛ/ into which it merged in the middle of the 20th century (Xu & Tang, 1988) and from which it separated from in the younger generation under Mandarin Chinese influence (Gu, 2004). It remains unclear whether such movement at the second half of the vowel serves to maximize the distinction between vowels with close positions at the phonemic level or it is just the phonetic nature of these vowels. In accordance with their movement, /ɤ, ɔ, e/ may also be phonetically transcribed as [^{ə}ɤ], [^{o}ɔ] and [e^{i}]. The downward and forward movement of /oʔ/ can be seen together with the movement of the other two high glottal-ended vowels /ɪʔ/ and /ʏʔ/, which shows the further centralization of glottal-ended vowels.

There is another consideration for transcribing /ɔ/ as [^{o}ɔ], which at the same time accounts

for the overlap of /u/ and /o/. As depicted in Fig.7, the vowel space of /u/ and /o/ was highly overlapping, hence the difficulty in distinguishing these two with duration, F1 and F2. Meanwhile, their phonotactic rules are not complementary. /u/ can be preceded by all consonants but palatals, i.e., /ɲ, tɕ, tɕ^{h}, dʑ, ɕ, ʑ/. /o/ cannot be preceded by palatals either. But there are more combinations missing since /o/ cannot be preceded by /f, v, t, t^{h}, d, l, g/, stemmed from either principled constraints or accidental constraints. Therefore, it is plausible to suggest that the difference lies in F3, which was not included in our study due to practical issues in measuring F3. We tentatively suggest that /u/ is actually unrounded [ɯ] while /o/ is actually rounded [u] based on the fact that the F2 of /u/ is slightly higher than that of /o/, although this needs to be confirmed by further studies. In this way, the high-back rounded /o/ is missing, leaving enough space for the vertical movement of /ɔ/.

The overlap or partial merger of the glottal-ended /əʔ/ and /ʌʔ/ has been documented with the generation born after 1965 in Xu and Tang (1988). From our acoustic data, these two phonemes are not (yet) fully merged in the generation born after 1995. However, it may be more accurate to transcribe /ʌʔ/ as /ɐʔ/ instead because it is more centralized as a glottal-ended vowel and its height is very close to /əʔ/ which lies in the middle. Fig. 9 shows the vowel space with data at both the midpoint and data at the upper quartile with transcriptions we proposed. They can be categorized into four groups following their distribution: (1) seven high front vowels /i, y, e, ɛ, ø, ɪʔ, ʏʔ/; (2) three middle vowels /ɤ, əʔ, ɐʔ/; (3) one low vowel /ɑ/; and (4) four high back vowels /ɯ, u, ɔ, oʔ/. More than two-thirds of the monophthongs gather in the upper half, and the feature of rounding is common with at least four round-unround pairs.

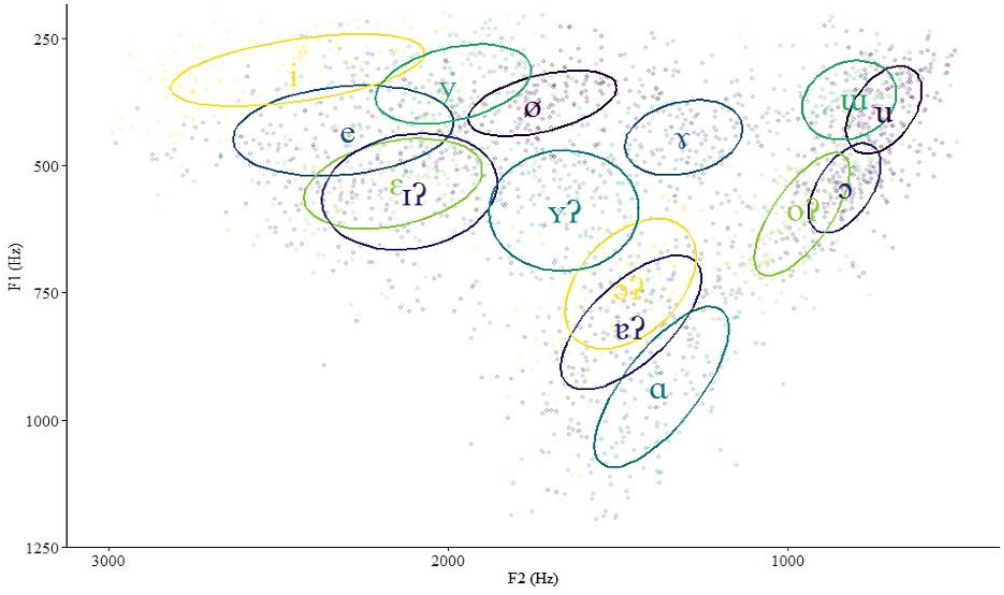


Fig. 9. Vowel space in Shanghai Chinese plotted with acoustic data from native speakers born after 1995 (n = 23). Each ellipse includes 50% of the data points.

Tone register turned out to be of importance in vowel height, which is illustrated by its interaction with syllable structure and vowel highness labels at both midpoint and upper quartile of vowels together with its main effect and interaction with gender at the upper quartile. Low-register vowels in general have a lower vowel height, measured with higher F1 values.

Nonetheless, its effect only occurs in some vowels, including the glottal-ended high vowels /ɪʔ, ʏʔ, oʔ/ and the open-ended low vowel /ɑ/ and cancels out the gender difference to some extent. What this can be associated to is that prominence increases the F1 values of glottal-ended vowels in general as observed by Chen (2008), who also suggested that the open-ended vowels remain stable spectral characteristics and are thus insensitive to prominence. It seems that the lower *f*0 functions differently than prominence. The effect of lower *f*0, howbeit, is unlikely to be a consequence of having phonologically voiced onsets; otherwise, we could expect a more pronounced effect at the midpoint than the 75% point of the vowel. The underlying articulatory mechanism remains unknown and deserves further research. In terms of vowel backness, the gender difference is more pronounced for front vowels than for back vowels.

4.3 Phonation type

No significant difference was found in breathiness between high- and low-register tones as measured by the difference between the first two harmonic amplitudes within the generation born after 1995. The 95% confidence interval of the effect size of tone register was between –1.84 to 1.52 dB/Hz, which was very close to 0. It might imply that the younger generation no longer produce low-register tones systematically with breathiness, in line with previous studies comparing the voice quality between the older and younger generations (Gao, 2016; Gao & Hallé, 2017; Tian & Kuang, 2016). Here we offer two possible accounts for the disappearance of systematic breathiness in production. First, as proposed by Gao et al. (2019), young speakers of Shanghai Chinese are mightily influenced by Mandarin Chinese in which the low tone has no breathiness cue and is not associated with the voicing status of the consonantal onset. In addition, there is a pronounced difference in relative pitch value between T2 and T3 albeit their similar pitch contour pattern. Therefore, breathiness may serve as a redundant cue in production. This account is also evidenced by the fact that the breathiness was already weaker in Shanghai Chinese in the older generation than in other subvariants of Wu (Gao & Hallé, 2017). Second, the breathiness might be canceled out by the creakiness in the low register which theoretically has a lower spectral tilt than the modal phonation while the breathy phonation has the highest amongst these three phonation types (Stevens, 1977).

Nevertheless, we did find an effect of syllable structure and its positive interaction with gender on the H1–H2 difference. These findings tentatively suggest that the voice quality of the two vowel sets is also dissimilar in the sense that open-ended vowels have a higher spectral tilt and less constricted voice quality than glottal-ended vowels, and the dissimilarity is more pronounced for female native speakers. The effect of syllable structure found in our study is in line with Gao and Kuang's (2022) production study of T3 and T5 contrast pairs showing that the main acoustic cues for checked tones (i.e., glottal-ended vowels) include a tenser phonation with greater vocal constriction and a slightly higher *f*0. More universally, our findings might potentially coincide with Maddieson and Ladefoged's (1985) study on four minority languages in China in which vowels with longer duration had more airflow coming out as measured by high positive *f*0–H2 difference. At this stage, it remains unclear whether and how voice quality is associated with duration and/or vowel space.

Another factor that might be relevant is the creakiness. Our results showed that low-register vowels are more likely to be produced with creakiness, realized as extra-low or

irregular *f*0. They are hence in line with studies on other languages like Mandarin (e.g., Kuang, 2017; Huang, 2020), Cantonese (e.g., Yu & Lam, 2014; Zhang & Kirby, 2020) and Vietnamese (Brunelle, 2009), in which creakiness is closely related to extra-low *f*0 (Huang, 2020). The creaky voice also occurs more often in open-ended vowels than glottal-ended vowels. More specifically, the difference holds between T3 and T2. Combining the creakiness with H1–H2 difference, open-ended vowels tend to have both less tense phonation and more creakiness. The latter can be explained by the fact that a substantial part of T3, the longest lexical tones, lies in the low *f*0 range and that T3 is in general the lowest among all five lexical tones in Shanghai Chinese (Chen & Gussenhoven, 2015; Gao & Hallé, 2017).

4.4 the LSTM network

Although our four-layer LSTM network trained with data from a wide range of age could not work properly to predict the vowel shift tendency over time for its mediocre F-score, we still found some patterns as seized by this recurrent neural network and revealed by its prediction based on the test set despite its relatively small size. First of all, open-ended and glottal-ended vowels are usually easily distinguished by the model because of the duration feature. Second, within each vowel set, F1 and F2 values play an important role in categorizing the vowel, as evidenced by the fact that most false predictions categorize the target vowel as a vowel with a close or overlapping space in the oral cavity. These two findings are in compliance with those from the statistical analysis. The model performed poorly in predicting vowels /ɪʔ, əʔ, y, e, ɯ/ in the sense that /əʔ/ was mistaken as /ʏʔ, ɐʔ, oʔ/, /y/ as /ø/, /e/ as /i, y, ɛ, ø/ and /ɪʔ, ɯ/ were easily mistaken as /ʏʔ, u/ but not the other way around. The asymmetric patterns can be resulted from two possible reasons: (1) The test data set is small; (2) The space and movement in the second half of the vowel has an impact on the prediction. We tentatively assume that both reasons hold in our case. In other words, it seems that the LSTM model may have (partially) seized the vowel space and movement entailed in the features. For instance, the height of /ʏʔ/ descended in the second half due to its centralization, making it closer to the position of /əʔ/. Similarly, /e/ moves upwards (and maybe forwards) in the second half, as illustrated in the statistical analysis. To figure out the underlying reason in this asymmetry, more acoustic data are needed for testing.

On the other hand, following the errors often made by our LSTM neural network, it can be expected that these vowels are also confused in real-life speech perception. The merger of /əʔ/ and /ɐʔ/ has been documented in Xu and Tang (1988), while it is the phonotactic rules that prevent /əʔ/ and /ʏʔ/ from being mistaken regarding their complementary distribution, except that both cannot be preceded by /ʑ/ and that both can be preceded by /ɦ/ or a zero onset. /y/ and /ø/ cannot exist in the same context either. The former can only exist after /l/ and palatals /ȵ, tɕ, tɕʰ, dʑ, ɕ, ʑ/, whist the latter cannot be preceded by these palatals. The fact that /e/ is easily mistaken as its surrounding vowels, indicates the instability of this relatively "new" phoneme as it diverged again among the generation born after 1965. The case of /ɪʔ/ mistaken as /ʏʔ/, /əʔ/ as /oʔ/ and /ɯ/ mistaken as /u/ cannot be attributed to the phonotactic rules nor any documented vowel shift. One possible account is that they differ in roundness, which is mainly measured with F3, yet included in neither the acoustic analysis nor as a feature in training the LSTM model. Note that the /əʔ/ − /oʔ/ roundness pair was not illustrated clearly from the

statistical analysis but only inferred from the confusion matrix of the LSTM network. In brief, we can tentatively conclude three situations in which vowels are hard to distinguish in the neural network. At the phonological level, they may be prohibited to exist in the same context by phonotactic rules. From the aspect of dynamic change, they may be in process of divergence or convergence, usually under the impact of extrinsic social factors. From a practical aspect regarding the design of the model, they may be distinguished by other features that are not included in the model.

More training data, particularly from the elder generation, are needed to improve the performance of this LSTM network in both precision and recall. An additional feature of F3 may largely improve the ability to distinguish /ɪʔ/ from /ʏʔ/, /əʔ/ from /oʔ/ and /ɯ/ from /u/ and may also contribute to the accuracy in distinguishing /i/ from /y/ and /e/ from /ø/ in Shanghai Chinese since these five pairs mainly differ in terms of roundness.

5. Conclusion

The present study on Shanghai Chinese contains a detailed report on the acoustic realization of vowels produced in isolation by the younger generation born after 1995 in terms of vowel duration, vowel space and voice quality and a report on the modeling of its diachronic change using an LSTM network with extra data from the elder generation aged between 40 and 60. The crucial findings are summarized as follows (also see Table 10):

A new monophthong system with minor adjustments is proposed: (1) high front /i, y, e, ɛ, ø, ɪʔ, ʏʔ/; (2) middle /ɤ, əʔ, ɐʔ/; (3) low /ɑ/; and (4) high back /ɯ, u, ɔ, oʔ/. More than two-thirds of the vowels gather in the upper half of the oral cavity, and the feature of roundness is quite common with five round-unround pairs (including the /əʔ/ − /oʔ/ pair according to the error analysis of the LSTM network). Lexical tones have a strong impact on vowels. In general, the tonal duration can be ranked as follows: T3 ≈ T2 > T5 > T4 (excluding T1). Open-ended vowels with unchecked tones (T1, T2 and T3) have quite different acoustic characteristics from those glottal-ended ones with checked tones (T4 and T5). The former are around 2.4 times longer than the latter when produced in isolation. The former also exhibit a more extensive space whilst the latter exhibit a more central vowel space. Low-register (T3 and T5) vowels in general seem to have lower height. Gender difference in F2 is more salient in front vowels. In terms of voice quality, no significant breathiness was detected at their beginning part in the generation born after 1995, which might result from its gradual disappearance among young native speakers under Mandarin Chinese influence. We also found a significant higher spectral tilt in open-ended vowels. Meanwhile, low-register vowels, especially those with T3, are more often produced with creaky voice realized with extra-low or irregular *f*0.

Apart from these findings from the statistical analysis, the four-layer LSTM network with duration and vowel space measures as features obtained an F-score of 0.6006 and a balanced performance between Type I and Type II errors. In spite of the mediocre performance of this bi-directional recurrent neural network which was likely a result of the small size of both training and test data, the detailed error analysis reveals three potential reasons for the wrong predictions: (1) The vowels are prohibited from existing in the same context by phonotactic rules, albeit their overlapping space; (2) The vowels are in process of convergence or

divergence, usually under the impact of extrinsic social factors, as the influence of Mandarin Chinese; and (3) Features as F3 are not included in the model.

Table 10. Findings in duration, vowel space and voice quality

		open-ended /i, y, e, ɛ, ø, ɤ, ɑ, ɯ, u, ɔ/	glottal-ended /ɪʔ, ʏʔ, əʔ, ɐʔ, oʔ/
number		10	5
duration	overall ratio	open-ended : glottal-ended ≈ 2.4 : 1 (in isolation)	
	register difference	low ≈ high (95% CI: −8.82 to 13.85 ms)	low > high
vowel space	distribution	more peripheral/extensive	more central
	F1	/ɑ/ low > high	/ɪʔ, ʏʔ, oʔ/ low > high
	F2	female–male difference more salient in front vowels	
voice quality	breathiness	higher spectral tilt	lower spectral tilt
		no significant difference between low and high registers	
	creakiness	more creakiness in T3	

References

Andruski, J. E. (2006). Tone clarity in mixed pitch/phonation-type tones. *Journal of Phonetics*, *34*(3), pp. 388–404.

Beguš, G. (2020). Generative adversarial phonology: Modeling unsupervised phonetic and phonological learning with neural networks. *Frontiers in Artificial Intelligence*, *3*, pp. 1–25.

Boersma, P., Benders, T., & Seinhorst, K. (2020). Neural network models for phonology and phonetics. *Journal of Language Modelling*, *8*(1), pp. 103–177.

Boersma, P., Chládková, K., & Benders, T. (2022). Phonological features emerge substance-freely from the phonetics and the morphology. *Canadian Journal of Linguistics*, *67*(4), pp. 611–669.

Boersma, P., & Weenink, D. (2023). Praat: doing phonetics by computer [software], Version 6.3.10, http://www.praat.org/.

Brunelle, M. (2009). Tone perception in Northern and Southern Vietnamese. *Journal of Phonetics*, *37*(1), pp. 79–96.

Cao, J., & Maddieson, I. (1992). An exploration of phonation types in Wu dialects of Chinese. *Journal of Phonetics*, *20*, pp. 77–92.

Chai, Y., & Garellek, M. (2022). On H1–H2 as an acoustic measure of linguistic phonation type. *The Journal of the Acoustical Society of America*, *152*(3), pp. 1856–1870.

Chai, Y., & Ye, S. (2022). Checked syllables, checked tones, and tone sandhi in Xiapu Min. *Languages*, *7*(1), 47.

Chao, Y.-R. (1928). *现代吴语的研究* [Studies in the modern Wu dialects]. Tsinghua College Research Institute.

Chen, Y. (2008). The acoustic realization of vowels of Shanghai Chinese. *Journal of Phonetics*,

36, pp. 629–648.
Chen, Y. (2011). How does phonology guide phonetics in segment – f0 interaction? *Journal of Phonetics*, *39*(4), pp. 612–625.
Chen, Y., & Gussenhoven, C. (2015). Shanghai Chinese. *Journal of the International Phonetic Association*, *45*(3), pp. 321–337.
Chen, Z. (2007). *上海市区话舒声阳调类合并的原因* [On the cause of the merger of the Yangping, Yangshang and Yangqu tones in the Shanghai dialect]. *方言* [Dialects], *29*(4), pp. 305–310.
Chollet, F., & others. (2015). Keras. GitHub. Retrieved from https://github.com/fchollet/keras.
Crystal, T. H., & House, A. S. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *Journal of the Acoustical Society of America*, *88*(1), pp. 101–112.
Delattre, P. (1962). Some factors of vowel duration and their cross-linguistic validity. *Journal of the Acoustical Society of America*, 34, pp. 1141-1143.
Duanmu, S. (1999). Metrical structure and tone: Evidence from Mandarin and Shanghai. J*ournal of East Asian Linguistics*, *8*, pp. 1–38.
Edkins, J. (1853). *A grammar of colloquial Chinese as exhibited in the Shanghai dialect*. Shanghai: Presbyterian Mission Press.
Esposito, C. (2012). An acoustic and electroglottographic study of White Hmong tone and phonation. *Journal of Phonetics*, 40, pp. 466–76.
Gao, J. (2016). Sociolinguistic motivations in sound change: On-going loss of low tone breathy voice in Shanghai Chinese. *Papers in Historical Phonology*, *1*, pp. 166–186.
Gao, J., & Hallé, P. (2017). Phonetic and phonological properties of tones in Shanghai Chinese. *Cahiers de Linguistique Asie Orientale*, *46 (1)*, pp. 1–31.
Gao, J., Hallé, P., & Draxler, C. (2019). Breathy voice and low-register: A case of trading relation in Shanghai Chinese tone perception? *Language and Speech*, *63*(2), pp. 1–26.
Gao, X., & Kuang, J. (2022). Phonation variation as a function of checked syllables and prosodic boundaries. *Languages*, *7*(3), 171.
Garellek, M., & Esposito, C. (2021). Phonetics of White Hmong vowel and tonal contrasts. *Journal of the International Phonetic Association*, *6*, pp. 1–20.
Gruber, J. F. (2011). An articulatory, acoustic, and auditory study of Burmese tone. Ph.D. Dissertation. Georgetown University.
Gu, Q. (2004). *最新派上海市区方言的语音调查* [New investigation of the sounds of Shanghainese city variety]. MA thesis, Shanghai Normal University.
Hillenbrand, J., Clark, M., & Nearey, T. (2001). Effects of consonant environment on vowel formant patterns. *Journal of the Acoustical Society of America*, *109*, pp. 748–763.
House, A. S. (1961). On vowel duration in English. *Journal of the Acoustical Society of America*, *33*, pp. 1174–1178.
Huang, Y. (2020). Different attributes of creaky voice distinctly affect Mandarin tonal perception. *Journal of the Acoustical Society of America*, *147*(3), pp. 1441–1458.
International Phonetic Association. (2015). The International Phonetic Alphabet (revised to 2015), http://www.internationalphoneticassociation.org/content/full-ipa-chart/.

Kuang, J. (2017). Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice. *Journal of the Acoustical Society of America*, *142(3)*, pp. 1693–1706.

Kuznetsova A, Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), pp. 1–26.

Lenth, R., Bolker, B., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Riebl, H., & Singmann, H. (2020). Emmeans: Estimated marginal means, Aka Least-Squares means. R package 1.8.6.

Ling, B.-J., & Liang, J. (2019). The nature of left- and right-dominant sandhi in Shanghai Chinese – Evidence from the effects of speech rate and focus conditions. *Lingua*, *218*, pp. 38–53.

Maddieson, I., & Ladefoged, P. (1985). "Tense" and "lax" in four minority languages of China. *Journal of Phonetics*, *13*(4), pp. 433–454.

Ohala, J., & Ewan, W. (1973). Speed of Pitch Change. *The Journal of the Acoustical Society of America*, *53*(1), pp. 345–345.

Qian, N. (2003). *上海语言发展史* [The history and development of the Shanghai Dialect]. Shanghai People's Press.

R Core Team. (2023). R: A language and environment for statistical computing [software], Version 4.3.0, http://www.r-project.org/.

Ren, N. (1992). *Phonation types and stop consonant distinctions: Shanghai Chinese*. Ph.D. Dissertation. The University of Connecticut.

Shen, Z. W., Wooters, C., & Wang, W. S.-Y. (1987). Closure duration in the classification of stops: A statistical analysis in a Festschrift for Ilse Lehiste. *Working Papers in Linguistics*, *35*, pp. 197–209.

Steinlen, A. (2002). *The influence of consonants on native and non-native vowel production*. Tubingen: Narr.

Stevens, K. (1977). Physics of laryngeal behavior and larynx modes. *Phonetica*, 34, pp. 264–279.

Stevens, K., & House, D. (1963). Perturbation of vowel articulations by consonantal context. *Journal of Speech Hearing Research*, *6*, pp. 111–128.

Svantesson, J. (1989). Shanghai vowels. *Working papers of Lund University*, *35*, pp. 191–202.

Tian, J., & Kuang, J. (2016). Revisiting the register contrast in Shanghainese. *Proceedings of the 5th Symposium on the Tonal Aspects of Languages*, pp. 147–151.

Vet, D. J. (2023). ED [software], Version 2023.08, http://www.fon.hum.uva.nl/dirk/ed.php.

Wong, P., & Chan, H.-Y. (2018). Acoustic characteristics of highly distinguishable Cantonese entering and non-entering tones. *The Journal of the Acoustical Society of America*, *143*(2), pp. 765–779.

Wu, F., & Kenstowicz, M. (2015). Duration reflexes of syllable structure in Mandarin. *Lingua*, *164*, pp. 87–99.

Xu, B., & Tang, Z. (1988). *上海市区方言志* [A description of Shanghainese spoken in the urban districts of the Shanghai City]. Shanghai Education Press.

Yu, K. M., & Lam, H. W. (2014). The role of creaky voice in Cantonese tonal perception. *Journal of the Acoustical Society of America*, *136*(3), pp. 1320–1333.

Zee, E. (1997). Duration and intensity as correlates of F0. *Journal of Phonetics*, 6, pp. 213–220.

Zee, E., & Maddieson, I. (1980). Tones and tone sandhi in Shanghai: Phonetic evidence and phonological analysis. *Glossa*, *14*, pp. 45–88.

Zhang, J., & Meng, Y. (2016). Structure-dependent tone sandhi in real and nonce disyllables in Shanghai Wu. *Journal of Phonetics*, *54*, pp. 169–201.

Zhang, Y., & Kirby, J. (2020). The role of F0 and phonation cues in Cantonese low tone perception. *Journal of the Acoustical Society of America*, *148*(1), EL40–EL45.

Appendix A. The profile of participants

Age	Gender	District of residence in Shanghai	Generation
23	female	Baoshan	young
23	female	Jiading	young
24	female	Baoshan	young
24	female	Baoshan	young
24	female	Jiading	young
24	female	Hongkou	young
24	female	Hongkou	young
24	female	Hongkou	young
24	female	Hongkou	young
24	female	Pudong (urban area)	young
24	female	Putuo	young
27	female	Hongkou	young
16	male	Hongkou	young
16	male	Hongkou	young
16	male	Yangpu	young
18	male	Changning	young
21	male	Hongkou	young
21	male	Putuo	young
24	male	Hongkou	young
24	male	Hongkou	young
24	male	Yangpu	young
25	male	Hongkou	young
26	male	Xuhui	young
40	female	Yangpu	elder
45	female	Pudong (urban area)	elder
47	female	Changning	elder
50	female	Hongkou	elder
56	male	Hongkou	elder
60	male	Hongkou	elder

Appendix B. Praat Scripts

1. Transforming the stereo sound to mono sound and removing silence

```
form stereo to mono
    comment Directory of sound files. Include the final "/".
    text sound_directory ../Vowel experiment/results/
    sentence Sound_file_extension .wav
endform

# Make a listing of all the sound files in a directory:
Create Strings as file list... list 'sound_directory$'*'sound_file_extension$'
numberOfFiles = Get number of strings

# Open each sound file in the directory:
for ifile from 1 to numberOfFiles
    select Strings list
    filename$ = Get string... ifile
    Read from file... 'sound_directory$''filename$'

    # get the name of the sound object:
    soundname$ = selected$ ("Sound", 1)

    # convert to mono
    select Sound 'soundname$'
    Convert to mono
    select Sound 'soundname$'
    Remove

    mononame$ = soundname$ + "_mono"
    select Sound 'mononame$'

    intensity = To Intensity: 50, 0.01, "no"
    textgrid = To TextGrid (silences): -25, 0.1, 0.075, "", "target"

    selectObject: sound, textgrid
    Extract intervals where: 1, "no", "contains", "target"

    # Save the selected intervals
    total_parts = numberOfSelected()
    for i to total_parts
        part[i] = selected(i)
```

```
    endfor

    # Rename
    sounding = Concatenate
    Rename: string$(sound)

    # Remove
    selectObject: intensity, textgrid
    # Including the extracted intervals
    for i to total_parts
        plusObject: part[i]
    endfor
    Remove

    selectObject: sounding
    # Simplify the name
    length = length(soundname$) - 22
    newname$ = mid$ (soundname$, 23, length)
    Rename... 'newname$'
    Write to WAV file... 'sound_directory$'/'newname$'.wav
    Remove
    select Sound 'mononame$'
    Remove

    # Delete the original stereo sound file
    filedelete 'sound_directory$'/'soundname$'.wav
endfor

select Strings list
Remove
```

2. Saving clipped sound files

```
form save_clipped_sound
    comment Directory of sound files. Include the final "/".
    text sound_directory ..Vowel experiment/results/
    sentence Sound_file_extension .wav
endform

# Make a listing of all the sound files in a directory:
Create Strings as file list... list 'sound_directory$'*'sound_file_extension$'
numberOfFiles = Get number of strings
```

```
# Write each sound file to the directory:
for ifile from 1 to numberOfFiles
    select Strings list
    soundname$ = Get string... ifile
    soundname$ = replace$ (soundname$, sound_file_extension$, "", 0)
    select Sound 'soundname$'
    Write to WAV file... 'sound_directory$'/'soundname$'.wav
    Remove
endfor

select Strings list
Remove
```

3. Getting duration, formant and harmonics

```
### The harmonics are obtained with three complete periods in which the given time
point falls in the middle.
### This script is written by Chengjia Ye in May, 2023. The formant analysis part
is based on the script "SemiAutoPitchAnalysis" by Daniel McCloy (drmccloy@uw.edu).
### If the output of H1 or H2 is -1, please check the spectrum manually. The
spectrum is not removed automatically in this case.

form Get duration formant and spectrum
    comment Directory of sound files. Include the final "/".
    text sound_directory ../Vowel experiment/results/
    sentence Sound_file_extension .wav
    comment Path of the resulting text file:
    text resultsfile ../resultsfile.txt

    comment Formant analysis parameters
    positive Time_step 0.005
    integer Maximum_number_of_formants 5
    positive Maximum_formant_(Hz) 5500
    positive Window_length_(s) 0.025
    real Preemphasis_from_(Hz) 50
endform

# Make a listing of all the sound files in a directory:
Create Strings as file list... list 'sound_directory$'*'sound_file_extension$'
numberOfFiles = Get number of strings

# Check if the result file exists:
if fileReadable (resultsfile$)
```

```
    pause The file 'resultsfile$' already exists! Do you want to overwrite it?
    filedelete 'resultsfile$'
endif

# Create a header row for the result file:
header$ = "Filename Duration    F1_midpoint F2_midpoint F1_0.75point    F2_0.75point
        ... H1  H2  H1-H2_difference    spectrum_starting   spectrum_ending
        ... f0_gap'newline$'"
fileappend "'resultsfile$'" 'header$'

# Open each sound file in the directory:
for ifile from 1 to numberOfFiles
    select Strings list
    filename$ = Get string... ifile
    Read from file... 'sound_directory$''filename$'

    # get the name of the sound object:
    soundname$ = selected$ ("Sound", 1)

    # get the duration of the sound object:
    duration_s = Get total duration
    duration = round(duration_s * 1000)
    midpoint = duration_s * 0.5
    upper_quartile = duration_s * 0.75

    # get the formant information of the sound object:
    To Formant (burg)... time_step maximum_number_of_formants maximum_formant
    ... window_length preemphasis_from

    f1_50 = Get value at time... 1 midpoint Hertz Linear
    f1_50 = round(f1_50)
    f2_50 = Get value at time... 2 midpoint Hertz Linear
    f2_50 = round(f2_50)
    f1_75 = Get value at time... 1 upper_quartile Hertz Linear
    f1_75 = round(f1_75)
    f2_75 = Get value at time... 2 upper_quartile Hertz Linear
    f2_75 = round(f2_75)
    select Formant 'soundname$'
    Remove

    # get the pointprocess
    select Sound 'soundname$'
    pitch = To Pitch (cc): 0.0, 75, 15, 0, 0.03, 0.45, 0.01, 0.35, 0.14, 600
```

```
select Sound 'soundname$'
plusObject: pitch
pointprocess = To PointProcess (cc)
removeObject: pitch

# get the beginning and end point of the given number of the period
selectObject: pointprocess
number_of_points = Get number of points
number_of_periods = Get number of periods: 0.0, 0.0, 0.0001, 0.02, 1.3
f0_gap = number_of_points - number_of_periods - 1

starting_time = Get time from index: 1
ending_time = Get time from index: 4

select Sound 'soundname$'
Extract part: starting_time, ending_time, "rectangular", 1.0, "yes"
To Spectrum: "no"

# remove the pointprocess file
if f0_gap < 1
    selectObject: pointprocess
    Remove
endif

# get the ending point in the entire sound
actual_start = round(starting_time/duration_s * 1000)/10
actual_end = round(ending_time/duration_s * 1000)/10

# get the H1 and H2 difference
partname$ = soundname$ + "_part"
select Spectrum 'partname$'
table = Tabulate: 1, 1, 0, 0, 0, 1
tablename$ = "Table " + partname$
b3 = object [tablename$, 3, 3]
b4 = object [tablename$, 4, 3]
b5 = object [tablename$, 5, 3]
b6 = object [tablename$, 6, 3]
b7 = object [tablename$, 7, 3]
b8 = object [tablename$, 8, 3]
if b4 > b3 and b4 > b5
    h1 = b4
else
    h1 = -1
```

```
    endif

    if b7 > b6 and b7 > b8
        h2 = b7
    else
        h2 = -1
    endif

    h1_h2 = h1 - h2
    selectObject: table
    Remove

    if h1 != -1 and h2 != -1
        select Spectrum 'partname$'
        Remove
        select Sound 'partname$'
        Remove
    endif

    # Save result to text file:
    resultline$ = "'soundname$''duration'  'f1_50' 'f2_50' 'f1_75' 'f2_75' 'h1'
        ... 'h2'    'h1_h2' 'actual_start'  'actual_end'    'f0_gap''newline$'"
    fileappend "'resultsfile$'" 'resultline$'

endfor
```