Perceptual clustering of high-pitched vowels in Chinese Yue Opera

Yixin Zhang, [1] Francis Nolan,[1] and Daniel Friedrichs[1,2,3]

[1] Phonetics Laboratory, Faculty of Modern and Medieval Languages and Linguistics, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DA, United Kingdom

[2] Department of Speech, Hearing and Phonetic Sciences, UCL, 2 Wakefield Street, London WC1N 1PF, United Kingdom

[3] Department of Computational Linguistics, University of Zurich, Andreasstrasse 15, 8050 Zurich, Switzerland

Abstract

Numerous studies on Western Opera singing have shown that listeners' vowel identification performance decreases with an increasing fundamental frequency (f_o). This study explores the intelligibility of high-pitched vowels in Yue Opera, the largest dialectal opera in China. Six long vowels (/i y e a o u/) were recorded by a professional female singer at ten f_os between 220 and 932 Hz, of which 700-ms nuclei with flat f_o contours and resonance trajectories were extracted as stimuli. In a within-subject design, sixteen phonetically trained listeners responded on a free-choice vowel quadrilateral (task 1) and in a two-alternative forced-choice task (task 2) to indicate which vowel was presented. Results show that vowels cluster in the perceptual space into three groups (/i y e/, /u o/, /a/) above 521 Hz and that listeners could identify vowels between but not within groups with high accuracy up to at least 932 Hz. Multidimensional scaling (MDS) of simulated auditory excitation patterns reveals highly differentiable spectral shapes between groups. These findings put into question whether previous results on Western Opera could be generalized to other forms of opera singing.

Key Words: Vowel Intelligibility, High-pitched Singing, Chinese Yue Opera

I. INTRODUCTION

It is often assumed that high-pitched singing is difficult to understand due to the loss of vowel intelligibility. The exploration of the loss has a rich history in Western Opera singing, and some literature has well summarized the relevant studies (e.g., Sundberg, 2013). As early as in 1885, von Helmholtz described an observation that the timbre of the vowel /u/ shifted towards /o/ when f_o in a male voice exceeded roughly 175 Hz (i.e., the musical note F3). A relatively recent study by Hollien et al. (2000) has confirmed this and shown that the identification of the sung vowels /i/ and /u/ shifted towards categories with F_1 just above the f_o in the sung stimulus vowels. Similar observations were reported in several other studies (for an overview, see Sundberg, 2013, p. 87), which all indicate that above a certain absolute f_o of approximately 523 Hz (i.e., the musical note C5) listeners' identification performance for all vowels but /a/ and /ɑ/ (which have the highest F_1 in normal speech) would successively decrease towards chance-level.

It is widely assumed among researchers from the field of singing that the aforementioned reduction in vocalic intelligibility is due to the sparse sampling of the vocal tract transfer function at high fundamental frequencies (f_o), which leads to a poor specification of the formants. As a soprano's vocal range reaches musical notes corresponding to f_o around 1 kHz (e.g., soprano C = 1046 Hz), the wide spacing of the harmonics makes it unlikely that typical formant frequency patterns can be found in the acoustic signal. This is particularly true for close vowels such as /i/ and /u/ that usually exhibit relatively low first formants (F_1), which would be exceeded by such high f_o.

However, studies outside Western Opera singing reported a non-uniform relationship between vowel intelligibility and f_o. For example, Smith and Scott (1980) found that the vowels /i ɪ e æ/ were identifiable (70% correct) up to an f_o of 880 Hz when they were produced in isolation by a soprano in a non-singing style with a raised larynx (i.e., shortened vocal tract). When asked to produce the same vowels in her typical soprano singing style at the corresponding musical note A5, the identification score dropped to 4%. The results of a relatively recent study by Nolan and Sykes (2015) have put these findings into question as it was shown that the vowels /i ɛ a ɑ ɔ u ə/ (produced in CV context with an initial lateral) all were perceived as or close to /ɑ/ at an f_o of 880 Hz (A5), although the soprano was asked to produce them in a non-singing style. On the contrary, Maurer and Landis (1996) demonstrated that the isolated vowels /i a u o/ (but not /e/) could be identified accurately by listeners between 497 and 873 Hz when they were produced by untrained children, women, and men at individually chosen f_o.

The contradicting results of these studies may be due to the uncontrolled secondary cues to vowel category perception (e.g., vowel duration, formant frequency movements, and co-articulation in the consonantal environment; for more information on this, see, for example, Strange et al., 1976, Lehiste and Peterson, 1961). Nevertheless, some studies that used excised vowels with a single duration and quasi-flat f_o contours and resonance trajectories still reported satisfying identification performance of the participants.

For example, Friedrichs et al. (2015a) found that the phonological function of the isolated steady-state vowels /i y e ø ɛ a u o/ can be maintained at f_os up to 880 Hz when they were tested in a listening test with only two response options. In a follow-up study investigating the influence of talker variability, Friedrichs et al. (2017) found that the cardinal vowels (point vowels) /i a u/ remained identifiable even up to 1046 Hz when they were tested in isolation and multiple response options were provided. In the same experiment, it was shown that listeners' identification performance decreased significantly for /y ɛ/ and dropped to chance for /e ø o/ within the range of 523–1046 Hz. Based on the analyses of auditory excitation pattern simulations, the authors proposed that the overall spectral shape of the cardinal vowels /i a u/ may be utilized by listeners as acoustic landmarks that aid vowel perception at high f_o. This assumption is supported by several studies that indicated that gross spectral shapes as represented by, for example, Mel Frequency Cepstral Coefficients (MFCCs) (Davis and Mermelstein, 1980) carry superior acoustic cues to vowel category identification than formants (e.g., Ito et al., 2001, and Zahorian and Jagharghi, 1993).

As many studies suggest that vowel identification is possible even when no typical formant patterns can be found in the acoustic signal (for a comprehensive overview, see Maurer 2016), it seems plausible that the reduction in vocalic intelligibility, especially the bias towards open vowels in high-pitched Western Opera, may largely be due to its special singing style. Joliveau et al. (2004) have demonstrated that Western Opera singers shift their first resonance (by opening their jaws and lips), and hence F_1, to the vicinity of f_o when they are singing at high pitches to gain vocal power. This so-called resonance or formant tuning may be beneficial when performing in large auditoria without microphones. However, such adjustments made to the articulation inevitably lead to changes in the acoustic patterns, which may explain the previously described migration of vowel category perception to those with higher F_1. Therefore, it is not only the listeners who are 'mishearing' but also the singers who are 'mispronouncing' the vowels in Western Opera.

Western Opera are not the only musical drama that became popular before the wide use of microphones, which may have played an important role in the evolution of contemporary singing styles. Various styles of musical drama also exist in China, in which the characters are played by specially trained singers. Similar to their Western counterparts, these Chinese Opera singers need to sing loudly while achieving a certain aesthetical norm. More importantly, many Chinese people also find Chinese Operas hard to understand. The naïve audience often attributes the unintelligibility of Chinese Operas to their special music styles, slow rhythm, and stylised languages, but whether changes in vowel qualities with changing f_o contribute to this has rarely been explored.

The only study on vowel intelligibility in Chinese Operas was carried out by Maurer et al. (2014), who examined the identifiability of vowels in Cantonese Opera singing by phonetically trained Cantonese speakers. The results showed high identification scores (>80% correct responses) for the vowels /i a ɔ u/ (but not /y œ/) in consonant-vowel (CV) or consonant-vowel-consonant (CVC) context. It is worth mentioning that the stimuli used in this study were extracted from a DVD of a famous female Cantonese Opera singer. Therefore, the vowels were not separated from the melody, and they could only roughly control the f_o levels. Some of these vowels might have co-occurred with musical notes that reflect the lyrics' lexical tones (for more information on tone and melody, see Wee, 2007 and You, 2006). In this way, the melody associated with the nine tones in Cantonese may narrow the lexical set and contribute to the identification by the native speakers. Thus, whether and why there is also a decrease in vocalic intelligibility as f_o increases in Chinese Operas requires empirical investigation under stricter conditions, namely, using vowels produced in isolation at strictly controlled f_os. The influence of the melody, tone as well as some secondary cues should be carefully controlled.

Another Chinese Opera style, which has not been studied in this context yet is Yue Opera. Unlike Western Opera and other Chinese Operas, it has the unique feature that all its characters, including all gender and ages, are played by females. This obviously requires a vast amount of control over phonation and articulation. The language used in Yue Opera is a stylized language specific to the use on stage. This language is based on the Wu dialect spoken in Shengzhou but also influenced by Mandarin during the development of Yue Opera (Qiu, 1995). Wu and Mandarin (and many other Chinese dialects) share the same logographic writing systems and similar syllabic structures. In both dialects, each character corresponds to a single morpheme and a syllable in the form of $C^{G}VC$ (C: consonant, G: glide, V: vowel or diphthong), and the onset and coda consonants are optional; namely, open syllables with a vowel or a diphthong only are allowed. However, the same character usually has different pronunciations in Wu and Mandarin dialects, and these two dialects have different vocalic, consonantal, and tonal inventories. According to previous studies (You, 2006; Huang, 2000; Qiu, 1999), the stage language of Yue Opera includes 13 groups of rhymes (i.e. *Dazhe*, ‘big rhymes’). The rhymes within the same group are considered as rhyming with each other, though they do not necessarily contain the same segments (e.g. /a/, /ia/, /ua/, /aʔ/, /iaʔ/, /uaʔ/ all belong to the same *Dazhe*). These 13 groups involve 20 long, short, nasalized, or dentalised vowels, with eight of them being long vowels (i.e., /i/ /y/ /e/ /u/ /o/ /ɔ/ /a/ /ɤ/). It is noteworthy that the male and female characters tend to realize some vowels slightly different from each other in Yue Opera, and the mid-close vowel is more commonly realized as /e/ by female characters but /ɛ/ when a male character is played (personal communication with Shuyang Sheng, the invited singer, and Weitao Mao, a well-known male character player and the vice-chancellor of the China Theatre Association).

In the present study, we recorded a professional female Yue Opera singer producing seven long isolated vowels (/i y e a ɔ o u/) in her singing style at ten f_0s between 220 Hz and 932 Hz by presenting her the corresponding morphemic characters containing the vowels only. The vowel /ɤ/ was not recorded because no morpheme corresponds to an open syllable with /ɤ/, namely, the very few morphemic characters containing /ɤ/ all have onset consonants that interfere with vowel quality. We conducted listening tests to compare the results with those from previous studies on Western Opera singing. To investigate the spectral properties underlying the listeners' identification process at high pitches, multidimensional scaling (MDS) was employed to geometrically model the changes in the perceptual space and simple versions of excitations patterns were analyzed that the vowels would be expected to generate in the auditory periphery.

II. METHODS

A. Participants

Fifteen phonetically trained listeners participated in the perceptual experiments (7 females, 8 males; mean age = 24.6, standard deviation = 3.5). They were all students at the University of Cambridge, and none of them reported any hearing impairments when asked before the experiment.

B. Stimuli and Apparatus

A professional female Yue Opera singer (age = 35) who received special training since school age was recorded in a noise-controlled room at the Phonetics Laboratory of the University of Cambridge using a MixPre-6 recorder and a Sennheiser M64 microphone with a K6 battery module. The sampling frequency of the recordings was 44100 Hz. She was asked to produce the vowels /i y u e o a ɔ/ in the Yue Opera style at ten f_os corresponding to musical notes between A4 to B$^\flat$6 without lexical tones (i.e., 220, 350, 440, 521, 659, 740, 784, 831, 880, and 932 Hz). Piano notes were presented as reference sounds to the singer via Sony MDR-Z7M2 headphones before each vowel production. She was asked to produce long and monotone vowels as accurately as possible while keeping a constant distance from the microphone of approximately 30 cm. The recordings were done twice to elicit more accurate stimuli, once by vowel (i.e., recording each vowel at all f_os before moving on to the next vowel) and once by f_o (i.e., recording all vowels at one f_o before moving on to the next f_o). The vowel recordings with the most accurate f_o realization were selected as stimuli. For reference purposes, spoken versions of the vowels in Yue Opera style were recorded at an f_o that was comfortable for the singer (mean f_o = 376.8 Hz, standard error = 25.93 Hz).

As the singer was unfamiliar with the international phonetic alphabet, she was presented with logographic characters corresponding to open syllables containing the target vowels. For each vowel, three different characters were presented to the singer to ensure the correct elicitation of each vowel. The characters used during the recordings were taken from Huang (2000), and the singer confirmed that the three characters in each group share the same vowel.

After the recording session, it was found that the singer diphthongized the vowel /ɔ/ into /ɔu/ throughout almost all f_os. This change may be due to the influence of Mandarin, the common language the singer used in conversational speech, in which the /ɔ/-carrying syllables are realized as /ɑʊ/. This diphthongization makes it impossible to investigate the categorical perception of /ɔ/ as a single vowel in the two perceptual tasks, so that /ɔ/ was dropped from the subsequent experiment and analysis. Only the recordings of the six long vowels /i y u e o a/ were used.

For each stimulus, 700-ms sound segments were extracted from the vowel centers. The excised sounds showed relatively flat f_o contours with a maximum deviation from the target f_o of 4 %. The sounds were normalized in Praat (Borsma & Weenink, 2021) to 75 dB SPL, and the onsets and offsets of the sounds were faded over 5ms by amplitude modulating the waveform with raised cosines. During the experiment, the output level was adjusted by listeners individually to a comfortable listening level.

C. Procedure

The perceptual experiment involved a guided transcription task and a two-alternative forced-choice task conducted successively through E-prime 2.0 (Psychology Software Tools, Pittsburgh, PA). The guided transcription task was chosen to investigate possible gradual changes in the vocalic intelligibility at different f_os, while the subsequent two-alternative forced-choice task allowed a more refined exploration of the categorical perception of different vowels. The participants could take a break as long as they wanted between the two tasks.

In the guided transcription task, the six vowels were presented at ten f_os in a pseudo-randomized order, resulting in 60 trials (6 vowels × 10 f_os). In each trial, the participants were presented with a figure representing the perceptual vowel space (including reference vowels, see FIG. 1) after receiving a vowel as an auditory stimulus. The perceptual space was presented as the vowel quadrilateral, in which the position of the vowels reflected a two-formant space. For instance, front rounded vowels were shown retracted from fully front.

The participants were asked to click at any point on the figure to indicate where they thought the vowel in the stimulus belonged to in the perceptual space. After the click, the screen would refresh automatically, signaling the start of the new trial, and the participants would hear the next stimulus simultaneously. The coordinates of their clicks were recorded. There was no time limit.

In the two-alternative forced-choice experiment, 300 trials were involved (6 intended vowels × 10 f_os × 5 noise vowels). In each trial, the participants were presented with an auditory stimulus and saw a screen that contained two horizontally arranged vowels out of the six, one of the two being the vowel intended by the singer. The left-right order of the vowel pairs, as well as the order of the auditory stimuli, was pseudo-randomized. The participants were asked to indicate whether it was the vowel on the right or the left they had heard by pressing two keys on the computer keyboard that were labeled beforehand by the investigator as ‘right’ or ‘left’. After the participants made their choice, they would hear the next stimulus automatically. There was no time limit, and the participants could only listen to a stimulus once.

D. Perceptual Data Analysis

To analyze the results of the guided transcription, we indexed the change in vowel quality by the distance between the coordinates of the participants' clicks and the coordinates of the intended reference vowels on the diagram of perceptual space (henceforth *Perceptual Distance*). Here, *Perceptual Distance* is not used to index whether the participants made a correct or incorrect response, but the perceptual changes, which might also reflect the potential changes in the singer's articulatory strategy.

We constructed several linear mixed effects (LME) models in R (R core team, 2020) using lmer in lme4 (Bates et al., 2015). We selected the optimal fixed structure by using stepwise comparisons from the most complex effect to the simplest and the random effects by the smallest Akaike Information Criterion (AIC). The final model has *Perceptual Distance* as the dependent variable, f_o and *Intended Vowel* as the fixed effects and *Participant* as the random effect.

Following Friedrichs' design (2015a), the participants' responses in the two-alternative forced-choice task were analyzed with the bias-free non-parametric sensitivity measure A′ according to Signal Detection Theory (Tanner and Swets, 1954; Stanislaw and Todorov, 1999; Pallier, 2002) in R (R Core Team, 2020). Signal Detection Theory applies to the situation in which participants are asked to determine which one of the two categories (i.e., which one of a vowel pair in our case) a stimulus belongs to. The task generates two measures of behavioral performance: the hit rate and the false alarm rate. In the present study, the response option of the lower F_1 (i.e., the closer vowel) was arbitrarily assigned to the signal (signal vowel), the other to the noise (noise vowel). Then, a *hit* (H) referred to when "the signal vowel was presented and chosen", a *miss* to when "the signal vowel was presented but not chosen", a *false alarm* (F) to when "the noise vowel was presented but not chosen" and a *correct rejection* to when "the noise vowel was presented and chosen". In studies using Signal Detection Theory, H and F are transformed into indices of sensitivity and bias based on statistical models like A′ and d′ (Pollack and Norman, 1964; Smith, 1995; Zhang and Mueller, 2005). Here, A′ rather than d′ was used because it is a non-parametric measure that can deal with situations when *hit* or *false alarm* rates are 0 or 1. In such instances, d′, the z-score difference between the signal and noise distribution (=Z (H) - Z(F)), is either -infinite or +infinite (Zhang and Mueller, 2005). A′ was calculated using the following formula (1) (Zhang and Mueller, 2005: 207):

$$(1)\ \ A' = \begin{cases} 0.75 + \dfrac{H-F}{4} - F(1-H) & \text{when } F \le 0.5 \le H \\ 0.75 + \dfrac{H-F}{4} - \dfrac{F}{4H} & \text{when } F \le H < 0.5 \\ 0.75 + \dfrac{H-F}{4} - \dfrac{1-H}{4(1-F)} & \text{when } 0.5 < F \le H \end{cases}$$

A′ ranges between 0 and 1, 1 indicating maximum performance and 0.5 indicating chance performance. The participants' response bias was indexed by B''_D, which correlates to the slope of the receiver operating characteristic function at the point of observation. B''_D is calculated as described in formula (2) (Pallier, 2002) and ranges from -1 (maximum bias to the noise vowel) and 1 (maximum bias to the signal vowel).

$$(2)\ B''_D = \frac{(1-H)\times(1-F) - H\times F}{(1-H)\times(1-F) + H\times F}$$

Since a very high A′ leads to meaningless B''_D as it is based on a small number of misses and false alarms (Stanislaw and Todorov, 1999; Zhang and Mueller, 2005), we only calculated B''_D of the vowel pairs with A′ values smaller than 0.7. We pooled over the participants (N = 15) to calculate A′ for each *Intended Vowel Pair* at each f_o as each vowel was only presented once to each participant.

E. Acoustic Analyses

Acoustic analyses were conducted to help to understand the perceptual results. Simple simulated auditory excitation patterns of the vowel stimuli were computed using a 200-channel linear gammatone filter bank. The bandwidths and centre frequencies were calculated according to the ERB formulae given by Glasberg and Moore (1990). For each filter channel, the rms level of the output wave was calculated and converted to dB. To account for the transmission properties of the middle ear, a frequency weighting based on measurements made by Puria et al. (1997) was applied.

Classical multidimensional scaling (MDS) analysis (Shepard, 1962a, b) of the simulated auditory excitation patterns was further employed to geometrically model vowel changes at higher f_os in the auditory perceptual space. MDS has been shown in previous studies (e.g., Iverson & Kuhl, 1995; Kewley-Port & Atal, 1989) to be a good technique to illustrate the perceptual similarity of vowels. Each vowel at each f_o was assigned to a point in a two-dimensional geometric space with distances in the MDS space linearly related to spectral distance. Hence, MDS can map the correspondence between perceptual and acoustic properties and show acoustic differences between and among phonetic categories across the different f_os.

III. Results

A. Perceptual Experiments

Guided Transcription Task. FIG. 1 shows that the basic shape of the transcribed vowel quadrilateral was maintained at all f_os, but the high front vowels /i y e/ as well as the high back vowels /u o/ started to cluster together from 740 Hz.

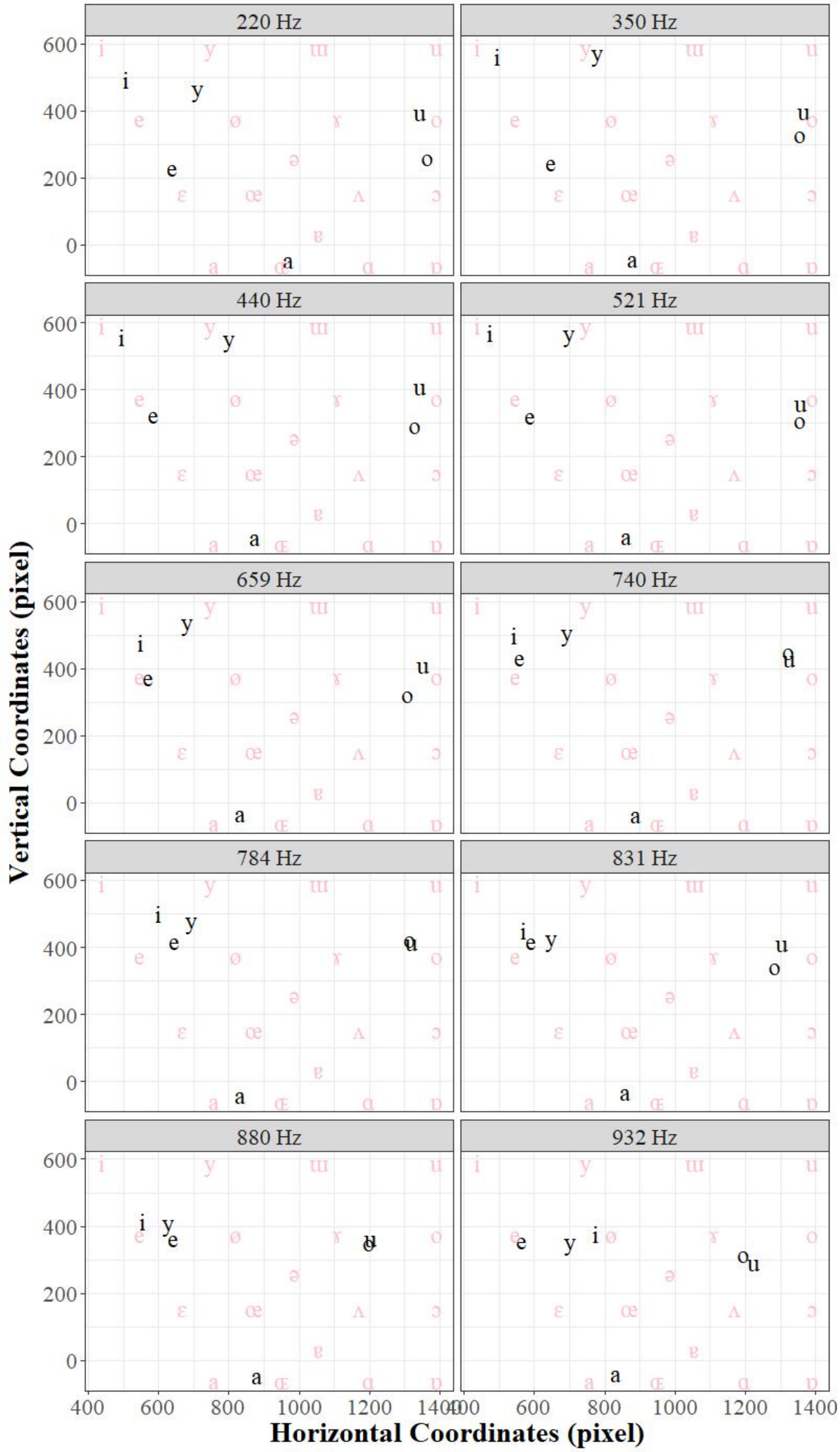
220 Hz
350 Hz
440 Hz
521 Hz
659 Hz
740 Hz
784 Hz
831 Hz
880 Hz
932 Hz
Vertical Coordinates (pixel)
Horizontal Coordinates (pixel)
0
200
400
600
400
600
800
1000
1200
1400

FIG. 1. (Color Required). Results of the guided transcription task for all f_os. Transcribed vowels are plotted in black at the averaged coordinates of the clicks. The reference vowels are shown in light red (i.e., the vowel quadrilateral shown to the participants). Note that the scales on the x and y-axes do not represent frequencies but numeric coordinates on a 20-inch screen (1600 × 900 pixels).

The clustering involved /i y u/ moving towards the categories with higher F_1, namely, towards [e ø o], and /e o/ moving up towards [i u] (henceforth, we describe phonemic vowels in Yue Opera in '/_/' and the vowels perceived by the listeners in '[_]'). From 831 Hz, the perceived categories of high vowels /i y u/ all shifted further towards the vowels with the next higher F_1. At the highest f_o, 932 Hz, /i/ was almost perceived as [e], and the perceived category of /u/ was close to [ʌ]. A closer examination of the *Perceptual Distance* (the distance between the average vowel placements in the guided transcription task and the relevant reference vowel on the quadrilateral) revealed that all the vowels except /a/ increased in mean *Perceptual Distance* from an f_o of 521 Hz (FIG. 2).

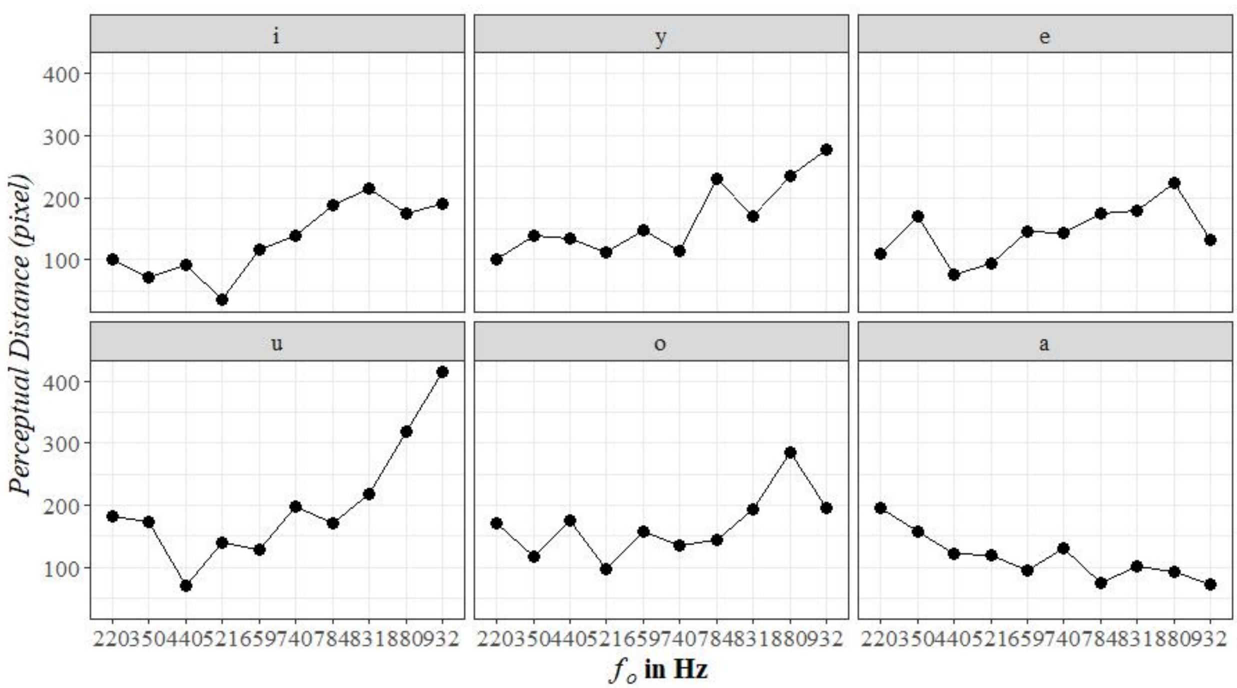

FIG. 2. Average *Perceptual Distance* for each vowel at all f_os. Note that the scale on the y-axes represents the distance in pixels between clicks and reference vowels on a 20-inch screen (1600 × 900 pixels).

The smallest average *Perceptual Distance* was often not found at the lowest f_o (220 Hz), but at the next higher f_os, which correspond to the singer's speaking f_o range. LME further revealed highly significant effects of *Intended Vowel*, f_o, as well as their interactions (TABLE I). The pairwise comparison (Tukey test) confirms that the differences in *Perceptual Distance* are significant (ps < 0.001) between the high f_os (880 and 932 Hz) and the relatively low f_os (440 and 521 Hz).

TABLE I. Results of the linear mixed-effects model on *Perceptual Distance* (Significance levels * =.05,** = .01, *** = .001)

Final Model	**Perceptual Distance ~ f_o + Intended Vowel +f_o: Intended Vowel + (1\Participant)**			
	SS	**df**	**F**	***p***
f_o	115.518	9	7.973	<.0001***
Intended Vowel	59.782	5	7.43	<.0001***
f_o: Intended Vowel	131.85	45	1.82	<.0001***

Two-alternative forced-choice task. A high identification accuracy was found throughout all f_os up to 932 Hz with median A′ above 0.75 (FIG. 3).

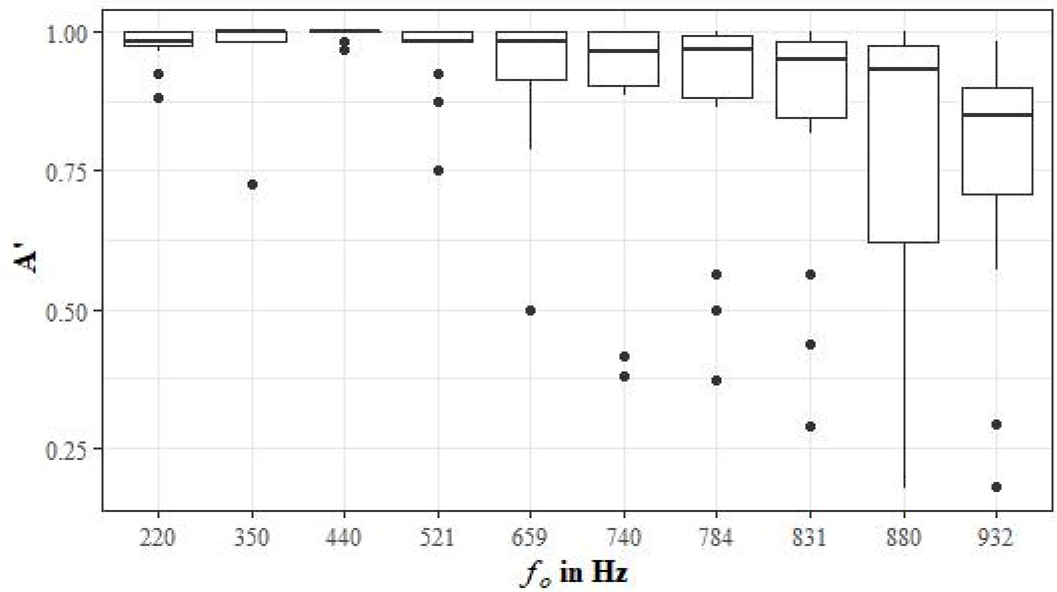

FIG. 3. Box plots showing the distributions of A′ (y-axis) for all vowel pairs that were tested at ten f_os between 220 and 932 Hz (x-axis). A′ of 0.5 represents chance level, and A′ of 1 represents maximum performance.

Vowel pairs involving the low vowel /a/ or pairs composed of front and back low vowels showed a stable and high identification accuracy across f_os up to 831 Hz (FIG. 4). At this f_o, A′ for these pairs ranged roughly between 0.75 and 1, except for /y-o/. In contrast, from an f_o of 659 and 740 Hz, respectively, A′ values for the pairs /u-o/ and /i-e/ dropped to chance level. The same observation was made for /i-y/ from 784 Hz upwards and for /y-e/ at 831 Hz before showing higher A′ again at the two highest f_os.

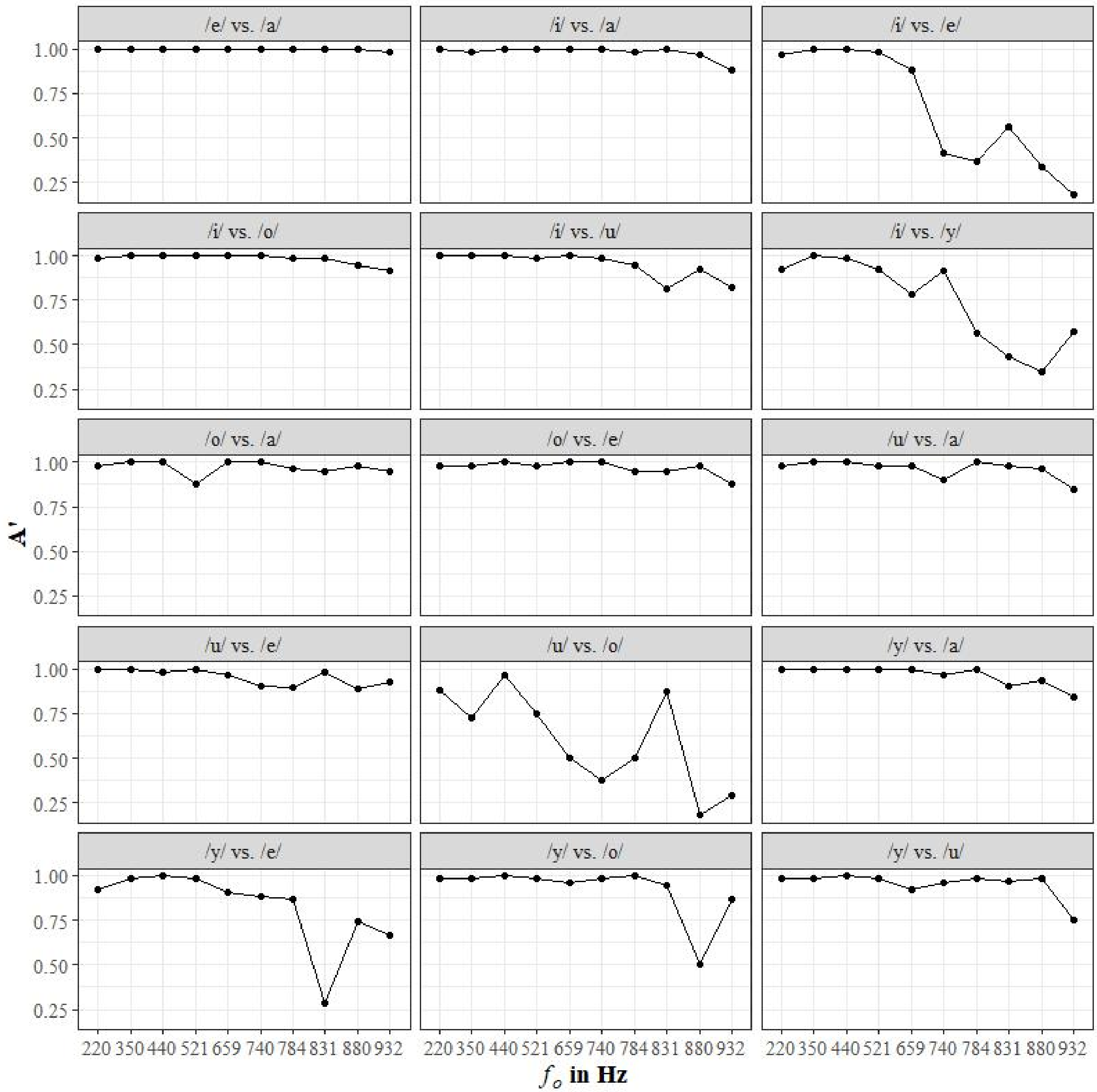

FIG. 4. A′ (y-axis) for each of the vowel pair contrasts at the ten investigated f_os (x-axis). A′ of 0.5 represents chance level, and A′ of 1 represents maximum performance.

Listener bias calculation is not meaningful when A′ is high as it is only based on a small number of misses or false alarms (Stanislaw and Todorov, 1999). We therefore only calculated B''_D for the vowel pairs /i-e/, /i-y/, /u-o/, /y-e/, and /y-o/ at the highest f_os from 659 Hz as A′ was smaller than 0.7 in these cases. No consistent bias was found for /i-e/, /i-y/, and /u-o/as results revealed both positive and negative B″D values with a high absolute value (e.g., -0.833 and 0.894). While no bias was found for the pair /y-e/ at any of the f_os, a strong bias towards /o/ was found for the pair /y-o/ at an f_o of 880 Hz, but at no other frequency. In other words, no consistent bias towards low vowels with higher F_1s could be observed.

B. Acoustics-derived auditory simulation

The results of the MDS analysis (FIG. 5) show the distribution of the stimulus vowels in a two-dimensional space derived from the spectral similarity of the simulated auditory excitation patterns. The spectral distances between the vowels at each f_o resembled to a high degree the perceptual results. Above 521 Hz, high front vowels (/y e/) started to cluster around /i/ while /u/ and /o/ started to cluster together and /a/ remained clearly separated from all other vowels. At the highest f_os, 880 and 932 Hz, the shape of the vowel quadrilateral was considerably less clear than in the perceptual space derived from the guided transcription.

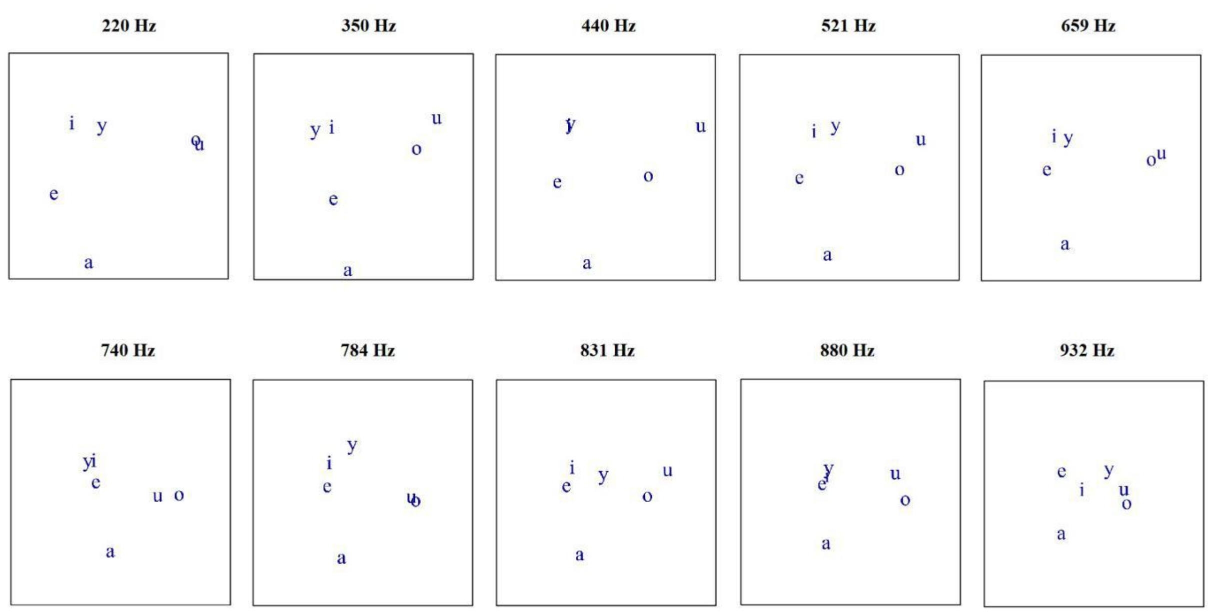

FIG. 5. MDS plots showing the auditory perceptual distance between the vowels used in this study throughout the f_os between 220 and 932 Hz. The differences between the vowels were

derived from simple versions of excitation patterns that they would be expected to generate in the auditory periphery.

A closer examination of the individual excitation patterns showed that despite the severe under-sampling of the vocal tract transfer function at very high f_os, the vowels /i u a/ still exhibited distinctive features up to at least 880 Hz (FIG. 6).

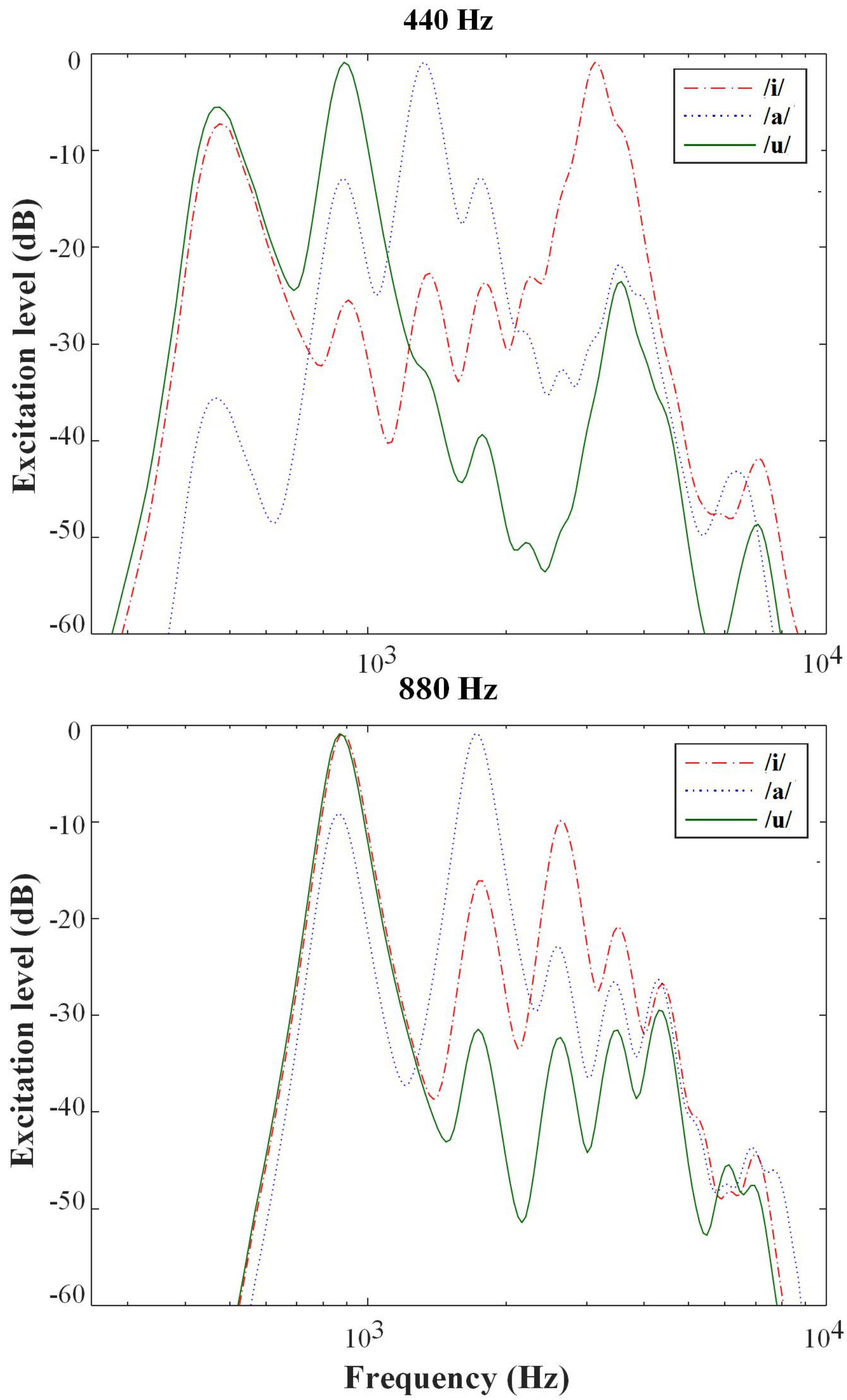

FIG. 6 (Color Required) Simulated auditory excitation patterns of the isolated vowels /i a u/ used in this study at an octave interval with f_os of 440 and 880 Hz. The excitation patterns reveal highly differentiable spectral representations at both f_os. At the higher f_o, the overall excitation

level in the frequency region above about 1.5 kHz can easily be distinguished. (The information in this figure may not be properly conveyed in black and white.)

The excitation patterns also showed a high degree of correspondence to the previously found confusion patterns in both perceptual tasks, namely, as f_o increased, the front high vowels tended to cluster together as well as the back vowels /u o/. For instance, the excitation patterns of /i y e/ and /u o/ at 880 Hz showed high within-group similarities (FIG. 7).

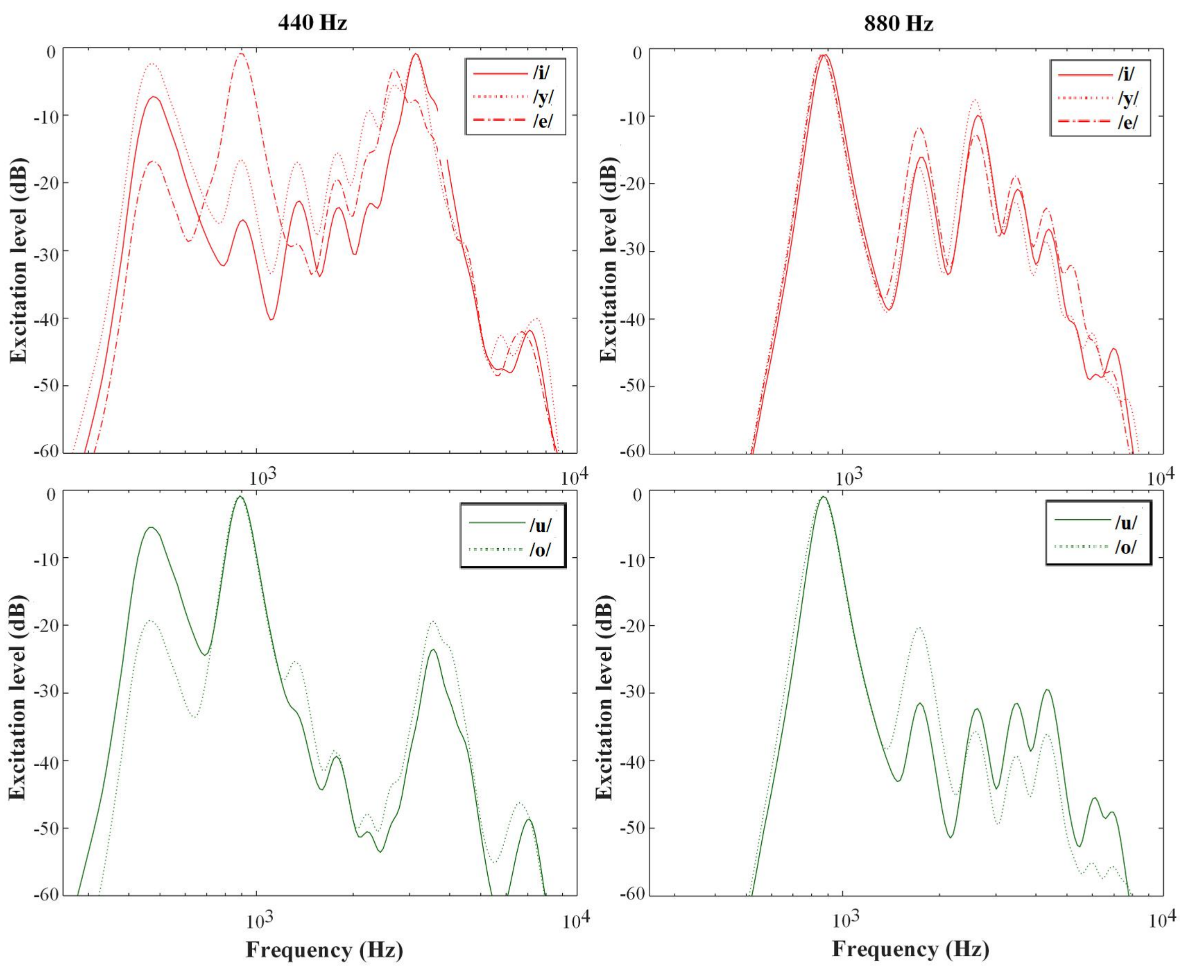

FIG. 7 (Color Required) Simulated auditory excitation patterns of the vowel groups /i y e/ and /u o/ used in this study. Both groups were found to cluster in the perceptual space at higher f_os. The

excitation patterns for 880 Hz reveal similar overall spectral shapes. (The information in this figure may not be properly conveyed in black and white.)

IV. Discussion

The results of the present study reveal a perceptual clustering phenomenon of the high front vowels /i y e/ and the high back vowels /u o/ at high f_os, with both clusters being highly differentiable from each other and from /a/. There was a considerable reduction in vowel distinctiveness in Yue Opera style singing within these clusters at higher pitches. However, the findings do not support the previously made assumption that listeners' identification would bias towards open vowels like [a] which many studies using vowels from Western Opera singers have suggested before (e.g., Nolan & Sykes, 2016; for an overview, see also Sundberg, 2013).

Furthermore, high vowels with low F_1 such as /i/ and /u/ were not always the first to lose their intelligibility, as found in some studies (Hollien et al., 2000, Howie and Delattre, 1962), but could even remain identifiable up to the highest f_o. The findings of the present study might, in fact, explain the results of Smith and Scott's research (1980) who reported good identification accuracy (70 % correct) for the vowels /i ɪ e æ/ when they were presented in a non-operatic style and in isolation at f_os around 880 Hz (i.e., A5). It seems likely that listeners could distinguish well between the two vowel pairs /i ɪ/ and /e æ/ but not always the vowels within the pair.

The guided transcription task used in this study revealed that the perceptual space resembles the basic shape of the vowel quadrilateral up to high registers in Yue Opera singing. However, the high front vowels /i y e/ started to cluster from 740 Hz and the high back vowels /u o/ from 659 Hz. Furthermore, towards the highest f_o investigated, the perceived categories of /i y/ shifted towards the lower categories [e, ø] and /u o/ towards [ɤ, ʌ]. The latter shift of /u o/ not only started from lower f_o but was also the most extensive observed in terms of perceptual distance. In contrast, the perceived category of /a/ remained accurate and stable across all the f_os.

The major confusions in the two-alternative forced-choice task were found between vowel pairs within a cluster, namely, /i-e/, /i-y/, /y-e/, and /u-o/ at 659 Hz and above. No significant confusions were observed between clusters, that is, for vowels from different clusters. As it is likely that no typical formant frequency distribution could be found in such vowel pairs at very high f_os, it seems likely that the overall spectral shape may carry enough acoustic information to distinguish between clusters, but not vowels within clusters.

Calculations of auditory excitation patterns and MDS analyses revealed apparent spectral differences between clusters and thus supported this hypothesis. Closer examination of the auditory excitation patterns revealed that the vowels constituting the observed clusters retain distinct spectral shapes, which kept them distinguishable from each other and /a/ throughout the f_o range investigated. However, at higher f_os, the vowels within each cluster exhibited very similar spectral shapes to one another, which may explain the decrease in listeners' identification performance in the two-alternative choice task for vowels within a cluster. These results indicate that highly differentiable overall spectral shapes (e.g., those representing [i a u]) can be used by listeners as acoustic landmarks to maintain some degree of vowel category perception at very high pitches. The calculations of the excitation patterns used in this study revealed distinct excitation levels in the frequency region above roughly 1.5 kHz for the vowels /i a u/, but highly similar levels for vowels within the clusters. Therefore, the present findings support the view that models of vowel perception based on formant peak patterns cannot provide such a full account of vowel perception as theories based on overall spectral shape (for a comprehensive review of several overall-spectral-shape models, see Kiefte et al., 2013).

This is also supported by the MDS analyses, which have shown that a triangular distribution of the six Yue Opera vowels could be observed up to about 880 Hz, though on a gradually reduced scale as an increasing f_o brought compression of the vowel space in both the dimensions of the MDS plots. The distance between /i/ and /u/ (and therefore *vowel frontness*) decreased as well as that between /i/ and /a/ (and therefore *vowel height*) (see FIG. 6). In contrast to Friedrichs et al. (2016), who observed an expansion in the front-back distinction when the vowel height dimension collapsed towards higher f_o of a Western Musical Theatre singer, the perceptual space containing the vowels produced by our Yue Opera singer did not show such compensation. This may either be due to the singer's personal habit or could also be because Yue Opera singers employ other mechanisms to protect the distinctiveness of sung vowels at high pitches – for instance, the association with melody and the tones embedded in the melody. A kinetic pitch on a vowel sweeps the transfer function with any available harmonic and therefore better reveals it than a static harmonic. More importantly, as previously mentioned, lexical tones embedded in the melody of an actual performance will also help lexical access by narrowing the candidates. Chinese theatre composition requires the composers and lyricists to follow the rule that the melody associated with each syllable should not conflict with the lexical tone in the beginning part, only allowing limited modification of the tone contour (Wee, 2007; Zhang, 1980: 91). Although in Mandarin popular song composition, this rule may not be followed strictly (for an overview on correspondence between lexical tone and sung melody, see Schellenberg and Gick, 2020: Table 1), theatre composition is much stricter on tone-melody harmony (Zhou and You, 1997: 190).

A mismatch between the results of the two perceptual tasks used in the present study is worth special attention. In the guided transcription, the two vowels that typically exhibited the lowest F_1, /i/ and /u/, are on average placed near [e] and [o] respectively on the response quadrilateral, but the vowels with typically the next higher F_1, /e/ and /o/, were also located in this perceptual vicinity (i.e., perceived correctly and placed near the most relevant reference vowels [e] and [o]) at 932 Hz. However, in the two-alternative forced-choice, no bias towards /e/ or /o/ was found for these vowels at high f_os. Instead, /e/ was often identified as /i/ at 880 Hz. When f_o exceeded 784 Hz, participants showed a bias towards [u] rather than [o], which would be more consistent with the findings of the transcription task. The acoustic analyses cannot fully explain these mismatched bias patterns. It may be that the participants were sensitive to the changes in vowel quality, but the categorical perception of vowels may not necessarily correspond completely to the perceived quality but was influenced as well by other factors like the task. Previous studies have indicated that different tasks do affect the listener's identification performance. For instance, participants performed better when they were presented with more meaningful response options (e.g., written words containing the target vowel vs. vowel letters), fewer response options, and a lower degree of talker variability (Friedrichs et al., 2017, 2015a, 2015b).

The results presented here, especially the identifiability of the high vowels (i.e., those with typically low F_1) at high f_o, and the much higher f_o at which identifiability started to decline compared to the studies on Western Opera (for an overview, see Sundberg, 2013), may also partly be driven by the features of Yue Opera. As there are no male singers in traditional Yue Opera, male, female, and even child characters in a single performance are all played by female singers. In order to distinguish between the gender and age of the different characters they are portraying, Yue Opera singers employ style-specific aesthetic and articulatory adjustments. For example, singers typically portray female characters with a more reduced mouth opening than male characters. This makes resonance tuning, as described in studies on Western Opera singing, very unlikely because it requires articulatory actions to increase mouth opening (i.e., opening the jaw, widening the lips). It seems plausible that tongue height and advancement and anatomical dimensions such as those of the pharynx might play a role in distinguishing between characters and maintaining intelligibility. To investigate this further and fully understand the correlation between character gender and vowel realisation, experiments with more Yue Opera singers performing in different gender and age groups are required (for further discussion on *the influence of gender and age on vowel quality*, see Maurer et al., 2015). This investigation may even be expanded to other Chinese Operas, which involve males playing females. Further research in this area may also be helpful to test whether vowel clustering can solely explain the relatively high intelligibility at high f_os or whether other factors contribute to this.

V. Conclusion

The present study on Yue Opera demonstrated that vowels clustered in the perceptual space into three groups (/i y e/, /u o/, /a/) at high f_o above about 521 Hz, and that listeners were able to identify vowels between but not within groups with high accuracy up to 932 Hz. The results, therefore, show that previous findings on vowel intelligibility in Western Opera may be style-specific and cannot be generalized to other forms of opera singing. The findings presented here furthermore support the view that the overall spectral shape provides a more robust cue than formant peak patterns for the perception of the high-pitched vowels. Further studies on articulatory strategies in high-pitched Yue Opera singing may be useful to fully understand the underlying mechanisms resulting in the perceptual clustering of vowels at high f_os.

ACKNOWLEGEMENTS

Thanks to Shuyang Sheng, the National-level Yue Opera performer in Shanghai Yue Opera House, who was recorded for this study. Further thanks go to Nick Clark, whose software was used to perform the gammatone filtering, as well as Stuart Rosen and Paul Iverson for their help with the auditory excitation patterns and MDS, respectively. The first author would also like to thank the CHINA Scholarship COUNCIL (CSC) and Cambridge Trust for their doctoral scholarship. The last author was supported by the Forschungskredit of the University of Zurich, Grant No. FK-18-077, and the Swiss National Science Foundation (SNSF), Grants No. P400PG_180693 and P2ZHP1_168375.

REFERENCES (BIBLIOGRAPHIC)

Bates D.M., Mächler M., Bolker B., and Walker S. (**2015**). "Fitting Linear Mixed-Effects Models Using lme4." Journal of Statistical Software, **67**(1), 1–48. doi:10.18637/jss.v067.i01.

Boersma, P., and Weenink, D. (**2021**). "Praat: Doing phonetics by computer [computer program] (version 6.0.39)," http://www.praat.org/ (Last viewed January 22, 2021).

Davis, S., and Mermelstein, P. (**1980**). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE transactions on acoustics, speech, and signal processing, **28**(4), 357–366.

E-Prime (Version 2.0.8.22) [Computer software]. Pittsburgh, PA: Psychology Software Tools.

Friedrichs, D., Maurer, D., Rosen, S., and Dellwo, V. (**2017**). "Vowel recognition at fundamental frequencies up to 1 kHz reveals point vowels as acoustic landmarks," J. Acoust. Soc. Am. **142**(2), 1025–1033.

Friedrichs, D., Rosen, S., Iverson, P., Maurer, D., and Dellwo, V. (**2016**). “Mapping vowel categories at high fundamental frequencies using multidimensional saling of cochlea-scaled spectra,” J. Acoust. Soc. Am. **140** (1), 3219.

Friedrichs, D., Maurer, D., and Dellwo, V. (**2015a**). “The phonological function of vowels is maintained at fundamental frequencies up to 880 Hz,” J. Acoust. Soc. Am. **138**(1), EL36–EL42.

Friedrichs, D., Maurer, D., Suter, H., and Dellwo, V. (**2015b**). “Vowel identification at high fundamental frequencies in minimal pairs,” in Proceedings of the 18th International Congress on Phonetic Science, paper number 0438, pp. 1–5.

Glasberg, B. R., and Moore, B. C. (**1990**). “Derivation of auditory filter shapes from notched-noise data,” Hearing research, **47**(1-2), 103–138.

Hollien, H., Mendes-Schwartz, A. P., and Nielsen, K. (**2000**). “Perceptual confusions of high-pitched sung vowels,” Journal of Voice, **14**(2), 287–298.

Howie, J., and Delattre, P. (**1962**). “An experimental study of the effect of pitch on the intelligibility of vowels,” Natl. Assoc. Teachers Singing Bull. **18**(4), 6–9.

Huang, W. (**2000**). Phonology of Yue Opera (Doctoral dissertation, Fudan University). [黄玮. (2000). 越剧音韵研究. (博士论文, 复旦大学)].

Ito, M., Tsuchida, J., and Yano, M. (**2001**). “On the effectiveness of whole spectral shape for vowel perception,” J. Acoust. Soc. Am. **110**(2), 1141–1149.

Iverson, P., and Kuhl, P. (**1995**). “Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling,” J. Acoust. Soc. Am. **97**(1), 553–562.

Joliveau, E., Smith, J., and Wolfe, J. (**2004**). “Tuning of vocal tract resonance by sopranos,” Nature, **427**(6970), 116.

Kewley-Port, D., and Atal, B. S. (**1989**). "Perceptual differences between vowels in a limited phonetic space," J. Acoust. Soc. Am. **85**, 1726–1740.

Lehiste, I., and Peterson, G. E. (**1961**). “Transitions, glides, and diphthongs,” J. Acoust. Soc. Am. **33**(3), 268–277.

Maurer, D. (**2016**). “Acoustics of the Vowel – Preliminaries,” (Peter Lang AG, International Academic Publishers, Bern, Switzerland).

Maurer, D., Suter, H., Friedrichs, D., and Dellwo, V. (**2015**). “Gender and age differences in vowel-related formant patterns: What happens if men, women, and children produce vowels on different and on similar F0?,” J. Acoust. Soc. Am. **137**(4), 2416.

Maurer, D., Mok, P., Friedrichs, D., and Dellwo, V. (**2014**). “Intelligibility of high-pitched vowel sounds in the singing and speaking of a female Cantonese Opera singer,” in *15th Annual Conference of International Speech Communication Association*, 2132–2133.

Maurer, D., and Landis, T. (**1996**). “Intelligibility and spectral differences in high-pitched vowels,” Folia phoniatrica et logopaedica, **48**(1), 1–10.

McKight, P. E., and Najab, J. (**2010**). “Kruskal-wallis test,” The corsini encyclopedia of psychology, **1**.

Nolan, F., and Sykes, H. (**2015**). “Vowel and consonant identification at high pitch: the acoustics of soprano unintelligibility,” *Proceedings of the 18th International Congress of Phonetic Sciences*, *Paper number 14*, 1–5.

Pallier, C. (**2002**). “Computing discriminability and bias with the R software,” URL: http://www.pallier.org/pdfs/aprime.pdf (Last viewed October 13, 2020)

Pollack, I., & Norman, D. A. (**1964**). A non-parametric analysis of recognition experiments. *Psychonomic science, 1*(1), 125-126.

Puria, S., Peake, W. T., & Rosowski, J. J. (**1997**). "Sound-pressure measurements in the cochlear vestibule of human-cadaver ears," J. Acoust. Soc. Am. **101**(5), 2754–2770.

Qiu, D. (**1999**). The Rhyme List of Yue Opera. *Theatre Lyrics* (4), 4.[裘达人. (1999). 越剧唱词音韵表. *戏文*(4), 4.]

Qiu, D. (**1995**). On the compilation of the Rhyme List of Yue Opera. *Theatre Lyrics* (1), 2.[裘达人. (1995). 关于修订《越剧唱词音韵表》的琐见. *戏文*(1), 2.]

R Core Team. (**2020**). "R: A language and environment for statistical computing [computer software] (version 3.1.3.)," R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Schellenberg, M., & Gick, B. (**2020**). Microtonal variation in sung Cantonese. *Phonetica, 77*(2), 83-106.

Shepard, R. N. (**1962a**). "The analysis of proximities: Multidimensional scaling with an unknown distance function. I.," Psychometrika **27**, 125–140.

Shepard, R. N. (**1962b**). "The analysis of proximities: Multidimensional scaling with an unknown distance function. II.," Psychometrika **27**, 219–246.

Smith, L. A., and Scott, B. L. (**1980**). "Increasing the intelligibility of sung vowels," J. Acoust. Soc. Am. **67**(5), 1795–1797.

Smith, W.D. (**1995**). Clarification of sensitivity measure A'. *Journal of Mathematical Psychology 39*, 82–89

Stanislaw, H., and Todorov, N. (**1999**). "Calculation of signal detection theory measures," Behav. Res. Methods, Instrum., Comput. **31**, 137–149.

Strange, W., Verbrugge, R. R., Shankweiler, D. P., and Edman, T. R. (**1976**). "Consonant environment specifies vowel identity," J. Acoust. Soc. Am. **60**, 213–224.

Sundberg, J., Lã, F. M., and Gill, B. P. (**2013**). “Formant tuning strategies in professional male opera singers,” Journal of Voice, **27**(3), 278–288.

Sundberg, J. (**2012**). “Perception of singing,” in Psychology of Music, 3rd ed., edited by D. Deutsch (Academic Press, London), pp. 69–106.

Tanner, W.P., Jr., & Swets, J.A. (**1954**). A decision-making theory of visual detection. *Psychological Review 61*, 401–409.

Wee, L. H. (**2007**). “Unraveling the relation between Mandarin tones and musical melody,” Journal of Chinese Linguistics, **35**(1), 128–144.

You, R. (**2006**). *Phonology of Chinese Operas*. The Commercial Press. [游汝杰. (2006). *地方戏曲音韵研究*. 商务印书馆.]

Zahorian, S. A., & Jagharghi, A. J. (**1993**). “Spectral-shape features versus formants as acoustic correlates for vowels,” J. Acoust. Soc. Am. **94**(4), 1966–1982.

Zhang, G. (**1980**). *The art of Chinese Opera*. China Drama Press. [张庚. (1980). *戏曲艺术论*. 中国戏剧出版社.]

Zhang, J., & Mueller, S. T. (**2005**). A note on ROC analysis and non-parametric estimate of sensitivity. Psychometrika, **70**(1), 203–212. doi:10.1007/s11336-003-1119-8

Zhou, Z., & You, R. (**1997**). *Dialects and Chinese Culture*. Shanghai People's Publishing House. [周振鹤&游汝杰. (1997). *方言和中国文化*.上海人民出版社.]