Vowel Modification (Aggiustamento) in Soprano Voices

May Pik Yu Chan[1] & Youngah Do[2*]

[1]University of Pennsylvania & [2]University of Hong Kong

Abstract

Singers convey meaning via both text and music. As sopranos balance tone quality and diction, vowel intelligibility is often compromised at high pitches. This study examines how sopranos modify their vowels against an increasing fundamental, and in turn how such vowel modification affects vowel intelligibility. We examine the vowel modification process of three non-central vowels in Cantonese ([a], [ɛ] and [ɔ]) using the spectral centroid. Acoustic results suggest that overall vowel modification is conditioned by vowel height in mid-ranges and by vowel frontness in higher ranges. In a following perception task, listeners identified and discriminated vowels at pitches spanning an octave from A4 (nominally 440 Hz) to G♯5 (nominally 831 Hz). Results showed that perceptual accuracy rates of the three vowels' match their acoustic patterns. The overall results suggest that vowels are not modified in a unified way in sopranos' voices, implying that research on sopranos' singing strategies should consider vocalic differences.

Keywords: Vowel modification, Sopranos, Acoustics, Perception, COG, Spectral centroid

* Corresponding author: Youngah Do (youngah@hku.hk)

Introduction

Sopranos of the Western classical tradition sing with high pitches (Callaghan, 2000) and they often need to convey lyrics when singing. This gives challenges to singers, as it is not easy to produce target vowels clearly at high pitches. One of the reasons is that classical singers need to consider both *timbre* and *voice projection*. Controlling singers' desired *timbre* across a wide pitch range is not easy to achieve without training, because the vocal range is divided into groups of adjacent pitches known as *registers* (Lewcock et al., 2001). When an untrained singer sings through registers, unplanned timbral changes may occur because pitches in different registers are produced in different ways (Lewcock et al., 2001). In addition to that, classical singers often sing without amplification technology while keeping themselves heard over the orchestra (Doscher, 1994). To do so, singers *project their voices* by controlling their resonance space. This is typically achieved via *formant tuning* or *resonance tuning*[1] (Doscher, 1994). When voice projection is achieved via desirable resonance, singers are said to have a *vocal ring* (Doscher, 1994).

Acoustically, the *vocal ring* may be attributed to resonance clusters around 3000 Hz, known as the *singer's formant* (Doscher, 1994). Resonance tuning occurs when manipulating the resonance space to achieve the vocal ring (Doscher, 1994). This cluster of higher resonances (R_3 and above) could be achieved via specific articulatory conditions: The cross-sectional area in the pharynx must be at least six times wider than the area of the larynx tube opening; the laryngeal sinus must be wide in relation to the larynx; and the sinus piriformis must also be wide (Sundberg, 1974). The singer's formant has been found in altos, tenors, and basses, where it tends to remain constant even with f_o or vowel changes, while resonance tuning is realized differently for sopranos. When sopranos sing in high pitches, there may not be harmonics

[1] Although terms like "formant clusters" and "formant tuning" has been used in the literature, the term "resonance" (R_1, R_2 … R_n) will be used in this study to refer to the intrinsic vocal tract resonance frequencies independent of the voice source. Notations of harmonics, resonances and formants and relevant measurements will follow what was outlined in Titze et al. (2015).

around 3000 Hz to form resonance clusters (Sundberg et al., 2013), and higher pitches could be projected on their own (Bozeman, 2013). To achieve the vocal ring, sopranos thus keep their R_1 close to the f_o, reinforcing resonance often via jaw opening (Sundberg et al., 2013). This is done until the end of the soprano's upper register (i.e. around or above 1000 Hz), when the R_1 could no longer be raised. At that point, sopranos will switch to the flageolet register (D. G. Miller, 2000). Apart from R_1, there are other resonance frequencies crucial to sopranos' singing, i.e., R_2. According to the results in Joliveau *et al.* (2004), where they measured the vocal tract resonances of trained sopranos via broadband acoustic excitation, R_1 is tuned by sopranos at high pitches, and R_2 is also raised. They also found that for some vowels, R_2 tends to cross over the $2f_o$ from 600 to 800 Hz with no R_2 tuning. Some other studies, however, found that some sopranos also employ R_2:$2f_o$ tuning in addition to R_1:f_o tuning (e.g., Henrich et al., 2011).

In vocal pedagogy, the act of adjusting vowel resonances at different pitches is known as *aggiustamento* (Callaghan, 2000). R. Miller (2000, p. 124) describes vowel modification as a process where vowels are subtly neutralized when the pitch rises. He argues that vowels should generally be modified as follows when approaching the upper *passaggio*: [i] to [ɪ]; [ɪ] to [e]; [e] to [ɛ]; [ɛ] to [ɑ]; [ɑ] to [ɔ]; [ɔ] to [o]; [o] to [ʊ]; and [ʊ] to [u] (R. Miller, 2000, p. 124). Acoustically speaking, this suggests that front vowels raise their R_1, while back vowels lower their R_1. It was also noted that singers should avoid neutralizing all the vowels towards to schwa immediately after the upper *passaggio* (R. Miller, 2000, p. 125). For example, after the upper *passaggio,* front vowels are modified backwards towards the neutral center, and back vowels may modify their vowel height to become more closed (R. Miller, 2000, p. 125). This suggests that the R_2 of front vowels would gradually drop, and the R_1 of back vowels would gradually rise after the second *passaggio*. Overall, vowel contrasts gradually get neutralized in the process of vowel modification.

As vowel determining resonances are modified in this way in singing, vowel intelligibility in high pitches has to be compromised accordingly. Since the focus of the current study is on vowel intelligibility in soprano's singing, we will review specifics of the previous studies on this topic. Smith and Scott (1980) compared the intelligibility of four vowels /i/, /ɪ/, /ɛ/ and /æ/ sung by a soprano at F4, A4, C♯5, F5, A5 and C♯6 in four conditions, viz: (1) operatic, (2) in a consonant-vowel-consonant (CVC) environment, (3) with a raised larynx, and (4) in a CVC environment and with a raised larynx. Condition (4) had the best identification results, and condition (1) had the least identification rates, suggesting an effect of consonantal context and f_o on vowel intelligibility. Gottfried and Chew (1986) compared the intelligibility of ten American English vowels sung by a countertenor in their full voice (130, 165, 220, 260, 330 Hz) and head voice (220, 260, 330, 440, 520 Hz) in the following four conditions: (1) CVC syllables sung in the full voice; or (2) CVC syllables sung in the head voice, (3) center 200-ms portions of the token (vowel only) sung in the full voice, and (4) center 200-ms portions of the token sung in the head voice. They found that vowels are less identifiable at higher f_os; and purely vocalic conditions, back vowels, and vowels sung in the chest voice were more easily misidentified (Gottfried & Chew, 1986). Benolken and Swanson (1990) also found that higher pitches lead to greater difficulty in sung American English vowels' recognition, and that a vowel may be misheard to have a higher "first formant" than intended. In Friedrichs and colleagues' (2015) study, a Swiss-German Musical Theatre singer sang vowel pairs in CVC conditions and produced the words as intelligibly as possible at f_os between 220 and 880 Hz. The results showed that vowel intelligibility is maintained in both purely vocalic and CVC conditions up to 880 Hz. Another recent study asked three singers (one soprano, one Musical-Theatre singer, and one actress) to produce eight vowels in isolation at 220, 330, 440, 523, 587, 659, 698, 784, 880, 988, and 1046 Hz. Listeners were presented with the vowel nuclei and 8 responses, and identified the vowels. The results showed that for German listeners, /i/, /a/, and

/u/ can act as acoustic landmarks up to 1000 Hz, while other vowels like /y/ and /ɛ/ become less intelligible at higher f_os (Friedrichs et al., 2017). Deme's (2017) study included sense and nonsense words produced in CVC contexts by a soprano at 175, 247, 349, 494, 698, and 988 Hz. The results showed that most vowels' intelligibility decreases gradually at high f_os in both sense and non-sense words. The results summarized so far suggest that vowel intelligibility rates are generally reduced at high pitches.

Despite several empirical reports, as summarized above, research on vowel intelligibility has rarely been paired with formal acoustic analyses. A main reason for this is that perception should concern the entire acoustic signal (including both the source and filter), the analysis of which is challenging for singing data. A common way to analyze resonances from an acoustic signal is linear prediction coding, which may be used to predict formants in speech (Deng & O'Shaughnessy, 2003). However, since harmonics inevitably space out at higher f_os, linear prediction analysis is not suitable when the f_o exceeds 350 Hz (Monsen & Engebretson, 1983), as shown in Figure 1. Hence, existing studies in singing often have not relied on the entire acoustic output. For example, resonances in soprano singers have been measured via external broadband acoustic excitation (Henrich et al., 2011; Joliveau et al., 2004). For this reason, studies on intrinsic resonance have not often been paired with vowel intelligibility, and how resonance tuning affects vowel intelligibility and the modification of vowel frontness/backness has been glossed over in most literature so far.

[insert Figure 1.]

Figure 1. Narrowband spectra (window length = 0.03 s) of the vowel [ɛ] spoken at 233 Hz (left), sung at 262 Hz (middle), and sung at 988 Hz (right). Figures were generated in Praat (Boersma & Weenink, 2018).

The current study uses the spectral centroid (henceforth SC) to infer how vowels modify over an f_0 continuum. SC has been often used to compare frontness of fricatives like /s/ and /ʃ/ (Kris & S. Greg, 1997; Nittrouer et al., 1989), where /s/ (more front) has a higher SC than /ʃ/ (more back). SC is suitable to compare /s/ and /ʃ/ because they differ in frontness but not height. Note that vowels have both height and frontness distinctions, so the SC values of vowel sounds typically lie between F_1 and F_2 (Chistovich & Lublinskaya, 1979). However, considering that sopranos tunes R_1 to f_0 (Henrich et al., 2011; Joliveau et al., 2004; Sundberg, 1975), vowels tend not to differ much in height at high pitch ranges. In addition, the f_0 will not pass through a resonance, and R_1 should tune close to f_0 and stay a little above f_0. Therefore, we can infer that, like fricatives, a SC measurement reflects vowel frontness at high f_0s.[2] In fact, the use of SC to measure vowels is not unprecedented (e.g., Chistovich & Lublinskaya, 1979; Delattre et al., 1952; Fox et al., 2011). The spectral integration hypothesis acknowledges that the F_1-F_2 model in speech assumes (a) F_1 and F_2 are always perceived as distinct peaks and (b) the role of higher formants in speech perception is ignored (Hayward, 2016). The SC effect thus argues that the perceived frequency peak is an integration of two formants, which is usually closer to the original peak that is stronger in amplitude (Chistovich & Lublinskaya, 1979).

The phenomenon of spectral averaging was first observed by Delattre, Liberman, Cooper and Gerstman (1952) who found that the vowels /u o ɔ ɒ ɑ a/ are identifiable when replaced with a "one-formant" equivalent. However, it was difficult to find one-formant equivalents for vowels like /æ ɛ e/, as their first two formants are further apart. For such cases, Chistovich et al. (e.g. 1979) proposed to look at the center-of-gravity (COG)[3], which hypothesizes that within a bandwidth of 3.5 Bark, listeners attend to changes in relative amplitude ratios between the

[2] Note that SC measurements are not as affected by the spacing out of harmonics at high pitches, unlike LPC measures, because it considers component frequencies' relative amplitudes against each other. Subsequently, both source and filter effects are reflected in the SC.

[3] This is identical to the spectral centroid (SC).

first two formants. This methodological choice was justified by some studies (eg., Fox et al., 2011), while exceptions exist (e.g., Assmann, 1991; Beddor & Hawkins, 1990).

Using SC measurements, the present study explores (a) the extent to which vowel intelligibility is compromised in sopranos' singing (b) and the nature of resonance tuning. To understand the modification tendencies of vowels differing in multiple features, vowels differing in both the height and backness dimensions are of interest. The present study explores vowel modifications of three Cantonese vowels, i.e., [a][4], [ɛ] and [ɔ], against an increasing fundamental frequency, and its effects on vowel perceptibility. We predict that the overall trend of the SC will rise, following a rising f_o and its harmonics. We also predict that the SC measurements for the three vowels are not uniform in low pitches and may deviate from an overall rising trend. This is because the SC reflects resonance-harmonic interactions. Whenever a resonance meets a harmonic, the harmonic gets boosted and the SC will fall or rise depending on whether R_1 or R_2 is closer to a harmonic. When R_1:f_o tuning occurs, the SC values of different sung vowels will start to converge to reflect a gradual change in vowel backness when the f_o continues to rise. As SC measurements will primarily reflect R_1:f_o tuning, which we predict all vowels to engage in, we do not anticipate clear vowel differences based on the SC signal. Instead, if certain vowels pattern together with each other, we infer the vowel properties by which the pattern is conditioned. For example, if [ɛ] and [ɔ] pattern together, we may attribute their pattern to their non-low property.

Regarding perceptibility, we expect that vowel intelligibility rates across participants will reflect the SC analyses of the production task. This means that vowel intelligibility rates would gradually decrease as the SC differences between the vowels decrease. We predict that both

[4] Note that the Cantonese vowel inventory is controversial. Specifically, the low vowel is sometimes transcribed as [a:] or [ɑ:], and as such, Hashimoto (1972) transcribes it as [A:]. It is also sometimes described as a low central vowel, and is transcribed as [a] (Bauer & Benedict, 1997; Matthews & Yip, 1994). In this paper, we will consider the low vowel as [a], though as we will see, our participants also produce the low vowel with varying degrees of backness in speech.

identification of vowels and potential discrimination of different vowel pairs will be above chance in lower fundamental frequencies, and around chance after resonance tuning occurs, due to decreasing SC differences between vowels. We also expect that any potential unexpected differences among vowels that are observed in the acoustic signal are reflected in perception. These predictions will be tested through a voice production experiment (Experiment 1) and a perception test (Experiment 2).

Experiment One

In the first experiment, sopranos sang nine CV combinations across a pitch range of around two octaves in an ascending and descending manner, resembling vocal warm-up exercises. Acoustic data was gathered and analyzed to track the vowel modification process.

Methods and materials

Participants

Six Cantonese native female speakers were recruited. All of them were sopranos who have received between four and ten years of classical voice training. They were in the age range between 20 and 22. Five of the participants were undergraduate students majoring in music with voice as their main instrument. One of the participants did not major in music but has obtained the Associate Performance Diploma from Trinity College London (ATCL) in voice, which is equivalent in standard to the first year of an undergraduate degree in musical performance. She has also been hired by private music centers as a voice teacher. Based on Bunch and Chapman's (2000) singer classification, four of the singers were considered to be category 6.2 "singing teacher[s]", and two of the singers were considered to be category 7.2 "full-time voice student[s]". All participants warmed up their voices before they started the experiment. All participants provided signed consent forms, and the experiment protocol was approved by the Human Research Ethics Committee at <the authors' institute>.

Stimuli

The vowels chosen for the study include three non-central vowels, i.e., [a], [ɛ], and [ɔ]. These vowels are also found in commonly sung languages in Western art songs (e.g. mélodies and Lieder) and operatic repertoires, such as Italian, French and German. By investigating commonly sung vowels, we expected to make more general implications, not necessarily restricted to Cantonese. The three vowels are contrastive phonemes in Cantonese, so participants were able to read stimuli written in Chinese characters and pronounce the intended vowel naturally. [i] and [u] were not included in the study as they are in complementary distribution in Cantonese open syllables (Kwan, 2014), so it would not be possible to create minimal-pair stimuli including [i] and [u] in the consonantal frame used for the rest of the vowels. Each vowel appeared in three CV contexts, with the consonants [p], [t] and [l], totaling up to nine CV combinations (3Cs x 3Vs). The three consonants have the place of articulation closer to the front of the mouth, which minimized coarticulation effects between dorsal or nasal consonants and vowel quality. All the nine CV combinations were pronounced in tone 1, i.e., high level tone. A phonetic level tone was chosen to avoid unnatural singing of steady pitches in a contour tone, as repertoires in complex tonal languages are rather rare. A high tone was chosen as opposed to mid-level (tone three) or low-level (tone six) tones because all nine CV combinations in high tone are real words in Cantonese. The CV combinations were 巴 [paː˥], 啤 [pɛː˥], 波 [pɔː˥], 咑 [taː˥], 爹 [tɛː˥], 多 [tɔː˥], 啦 [laː˥], 哩 [lɛː˥ and 囉 [lɔː˥]. One consonant was presented for each scalic passage, with the vowels being sung in the order of [a], [ɛ] and [ɔ]. The CV combinations were all sung across a pitch range of around two octaves, starting from B3 (nominally 247 Hz) to C6 (nominally 1047 Hz), rising by semitone. At every pitch, all three vowels were sung in the same breath group to encourage singers to keep the same level of jaw opening, minimizing vowel height differences. The tokens sung at C6 marked the turning point of the ascending and descending scale. This note was not repeated as it only

served as a transition, consequently we did not include tokens sung at C6 in the analysis. Tokens sung at B3 were also excluded as they were the first and last notes, vulnerable to prosodic boundary tones. This means that data from all 9 CV combinations were collected at 24 different musical notes, starting at C4, which is within a normal pitch range for female adults' speaking voice.

To keep pitch accuracy and to control vowel length, an audio file was provided to participants to sing along with. The audio file was generated using Version 2018.6 of Sibelius First (2006), where the notes were set to the default synthesized piano sound in Sibelius. The notes were set to three crotchets per bar, set at a tempo of 88 beats per minute. Between each pitch change, there was a bar rest where the audio file played the note that was just sung, the next note to be sung, and a crotchet rest. This allowed the participants to expect the next pitch. The words that participants should sing were presented in the form of a color-coded musical manuscript to guide the singers along the audio file, which was also generated using Sibelius First. The Chinese characters were written below the musical stave as lyrics. Only the notes ranging up to 349 Hz were notated fully as guidelines, because the sequence of words was repeated at every note.

For comparison, we also made speech recordings of all CV combinations from our participants. The speaking task was conducted after the singing task. The stimuli were presented on the screen with the carrier phrase "佢話_____喎" [kʰɵyɻ waːɬ ____ kwaːɬ] (3sg say ___ SFP (sentence final particle)), meaning "I guess he/she said____". The words before and after the target word were both kept at a level (non-contour) tone with no lexical pitch changes, to minimize tonal influences.

Procedure

Singing Task. All recordings took place at a sound-attenuated booth at the authors' institute. For this task, participants stood next to a Sennheiser MD 46 microphone on a microphone stand.

Another Sennheiser MKE2P-K Lavalier Condenser microphone was attached to the participant's collar. Two microphones were used as the volume of the participants' singing gets significantly louder at higher pitches. Two different microphones were located at different areas to ensure that the participant's voice was picked up through her entire pitch range. An additional purpose of having two microphones was to backup recordings in case of clipping. Recording levels of both microphones were adjusted during silent moments of the recording sessions when clipping was anticipated. For each participant, the recording that was not clipped during any part of the entire recording session was chosen. In cases where both recordings were not clipped, the recording with a general greater amplitude was chosen for analysis. Recordings were made at 44.1 kHz sample rate and 16-bit sample depth. Audio was presented using Samsung EHS64 Wired Stereo Earbuds and participants only used the earbuds on one ear of their own choice, so that they can hear their own voices during the experiment. Participants were first given the visual stimuli presented on paper and asked if they could read all the words out once; this was to ensure that the participants know the intended pronunciation of each character. A practice run using the stimulus set 巴 [paːɭ], 啤 [pɛːɭ], 波 [pɔːɭ] was sung once. This allowed the participant to get used to reading Chinese characters as stimuli, and to adjust the distance of the microphone if needed.

Six sets of stimuli with different consonant orders were used across the participants. For example, one participant sang in the order of [pa] [pɛ] and [pɔ] followed by [ta] [tɛ] and [tɔ] and at last [la] [lɛ] and [lɔ], whereas another participant sang in the order of [t], [p], and then [l] or [p], [l], [t], to balance the order of consonants which could potentially affect the following vowels. The vowels were always sung in the same order across all consonantal contexts and participants. Each scale was sung in an ascending-descending order, such that there were two values for each CV combination at each pitch.

Speaking Task. The speech recordings were taken at the same location as above once the participants took a short break after the singing task. The same microphones were used, and sample-rate and bit-depth were also kept constant. Participants were seated in front of a computer, with a monitor placed at eye level at the participant's most comfortable distance. The stimuli in the carrier phrase were presented in a randomized order using the software Psychopy (Peirce & MacAskill, 2018), with two repetitions for each phrase. A 500 ms gap was left between each stimulus to indicate that the coming phrase was to be a new stimulus in case of consecutive stimuli repetition.

Data processing

In total, 2,862 tokens were collected and analyzed with Praat Version 6.0.43 (Boersma & Weenink, 2018). After the exclusion of recordings at the lowest (B3, nominally 247 Hz) and the highest (C6, nominally 1047 Hz) frequencies, 2,700 tokens were left. SC measurements for all tokens were taken over the middle 50% of the vowel's duration, minimizing influences from consonantal formant transitions. The Praat function "Spectrum: Get centre of gravity..." was used to obtain the SC measurements from Praat directly, and the data analysis process was automated via a Praat script. The F_1 and F_2 values of the speech tokens were measured at the vowel's mid-point using Praat's LPC (burg) analysis (Andersen, 1974).

The raw amplitudes of the first two harmonics were extracted using VoiceSauce (Shue, 2010) to obtain the ratio of the raw amplitudes of the first two harmonics. The window size was set to 25 ms, the frame shift was set to 1 ms, and the number of periods for harmonic estimation was set to 3. The STRAIGHT algorithm (Kawahara et al., 2008) was chosen for f_0 measurement, which was used for parameter (harmonic) estimation. H1H2u measures (henceforth A_1-A_2) were extracted at every frame (every millisecond) within the middle 50% of the vowel's total duration.

The non-linear relation between f_o and SC was analyzed by fitting a Generalized Additive Mixed Model (GAMM) (Wood, 2017) to the data, using the mgcv package (Wood, 2019). The SC was analyzed using a regression model including a multivariate normal error term. A smooth function of the f_o for each participant-vowel combination was included. In other words, a different curve was fit for each combination of participant and vowel. Fixed effects included were consonant onsets (3 factor levels: [l], [p], [t], with [l] as the reference level), vowels (3 factor levels: [a], [ɛ] and [ɔ], with [a] as the reference level) and target f_o (numeric). Each participant was also assumed to have their own random intercept. An improper normal prior distribution with the variance matrix being the pseudoinverse of the smoothing matrix was assumed for the purpose of credible interval calculation (Wood, 2017, p. 239). For the consonants, the modeling results showed that both [p] onset (β = -14.75, SE = 4.76, t = -3.10, Pr(>|t|) = 0.002) and [t] onset (β = -29.43, SE = 4.76, t = -6.18, Pr(>|t|) = 7.68e-10) were significantly different from the reference level [l]. For the vowels, both [ɛ] (β = -76.19, SE = 4.76, t = -15.99, Pr(>|t|) <2e-16) and [ɔ] (β = -82.96, SE = 4.76, t = -17.41, Pr(>|t|) <2e-16) were significantly different from the reference vowel [a], and a significant fixed effect of target frequency was also found (β = 1.54, SE = 0.11, t = 13.35, Pr(>|t|) <2e-16). The modelled output, which was generated from the raw data, will be referred to as "estimates" below.

Results

To show the SC against f_o change, the estimated mean plots of the SC are provided in Figure 2. 95% credible intervals for the modelled estimates are represented by the error bars. The credible intervals were calculated using the results of the GAMM. Overall, all participants showed the same trend: the vowels' mean SC were similar at the higher end of the pitch range, but SC across vowels were not uniform at lower pitches. Such pattern was predicted, since the SC at lower pitches should reflect both R_1 and R_2 values which varies across vowels. From

around 600 Hz to 700 Hz f_o (around D#5 to F#5), the SC of the three vowels started to converge and rose in a steady slope. Apart from participant 6, the SC stayed close to f_o, which implies that the singers' f_o was reinforced by a resonance, which was most likely R_1. Therefore, we inferred that R_1:f_o tuning has begun when the SC started to rise steadily. [ɛ] tended to start to modify first, followed by [ɔ] and [a].

[insert Figure 2.]

FIG. 2. Estimated mean SC by participant and vowel. x-axis: target fundamental frequency in Hz; y-axis: modelled mean spectral centroid in Hz; error bars: 95% credible intervals (calculated using the results of the GAMM). The diagonal grey lines show f_o to $3f_o$. The participants' average speech F_1s are included as vertical lines.

Note also that there was a gap between the participant's mean spoken R_1 and the start of R_1:f_o tuning. Figure 2 shows that the mean F_1 in speech for [ɛ] and [ɔ] was lower than the start of R_1:f_o tuning, and conversely, the mean F_1 in speech for [a] was much higher because [a] is a low vowel. The F_1 and F_2 of the spoken vowels were also provided in figure 3 for a comparison. Participants' spoken vowel [a] appeared to have varying degrees of frontness.

[Insert Figure 3.]

FIG. 3. F_1 and F_2 of the spoken vowels. x-axis: F2 in Hz; y-axis: F1 in Hz. y-intercept: minimum derivative of the SC between the 650-850 Hz range (~onset of R_1:f_o tuning)

Figure 4 compares the SC values of vowel pairs. In the lower half of the pitch range, the SC difference across the three vowels fluctuated as expected. By around 600 Hz to 700 Hz, the SC difference between [a] and [ɔ] tended to stay minimal, although differences involving [ɛ]

were still observed. Focusing on the upper pitch range, apart from participant 6, the 95% credible intervals of the vowel differences have already started to cross 0 by 740 Hz. After that, most of the vowels' SC differed minimally, although there is not sufficient evidence to conclude that the vowels' SC are different as reflected by the credible intervals. While it was predicted that the front vowel [ɛ] should always have a higher SC compared to [a] and [ɔ]; at higher frequencies, this only seemed to be the case for a subset of participants, i.e., participants 2, 3 and 5. The SC values of [a] or [ɔ] minus [ɛ] should result in a negative value if [ɛ] had a higher SC, and the red and green lines were only below the x-axis for these three participants. As the red and green lines stayed above 0 for participants 1 and 6, the SC of [ɛ] for these two participants were likely to be lower than that of [a] and [ɔ]

[insert Figure 4.]

FIG. 4. Estimated SC differences in vowel pairs. x-axis: target fundamental frequency in Hz; y-axis: modelled spectral centroid difference in Hz; error bars: 95% credible intervals.

We further extracted the raw A_1-A_2[5] values from the signal. Results were summarized in Figure 5. Around 400 to 800 Hz, the relative amplitude of A_2 for [a] and [ɔ] appeared to be stronger than that of [ɛ], suggesting that the R_2 for [a] and [ɔ] was closer to $2f_o$ than [ɛ] at that pitch range. The opposite was true for some participants once it reached 800 Hz. Note that, a negative A_1-A_2 value at low frequencies was not surprising because R_1 may have started at a higher frequency that was closer to $2f_o$ than to f_o. Around 400 Hz onwards, multiple singers showed a dip in A_1-A_2. This suggests that A_2 was stronger than A_1, which could occur if either R_1 or R_2 got close to $2f_o$. Since R_1 is the lower resonance, it was likely R_1 that caused the boost

[5] Note that amplitude was not strictly controlled in this experiment, as amplitude levels were adjusted during the recording process to prevent clipping, and participants may have moved closer to or further away from the microphone. Therefore, though it is still informative as it concerns a ratio, the overall trends in the plot are not fully reliable. For this reason, no statistical tests are applied here.

in A_2. As R_1 moved further away from $2f_o$, the A_1-A_2 values continued to raise. Sometimes a second weaker dip in A_1-A_2 could be observed, like participant 3's vowels between 500 and 600 Hz. This suggests that A_2 has been strengthened once again, which may have been caused by R_2 crossing over $2f_o$. As such, the A_1-A_2 value dipped and rose again. This pattern was not robust across all participants and vowels because R_1:f_o tuning may also have started at a similar pitch range, which would have raised A_1-A_2. If R_2:$2f_o$ tuning occurred simultaneously, the amplitudes of both A_1 and A_2 would have been strengthened, and A_1-A_2 would not have been affected much. After around 600 Hz, most participants' A_1-A_2 value continued to rise, suggesting strong R_1:f_o tuning effects.

[Insert Figure 5]

FIG. 5. A_1-A_2 by participant. x-axis: measured F0 in Hz; y-axis: difference in amplitude between the first and second harmonics.

Discussion

We infer that R_1 was likely tuned to f_o when the SC increased gradually. When R_1 and R_2 were modified simultaneously, the combined effect caused the SC to fluctuate. When R_1 was tuned to f_o, R_1 rose by about the same amount as the fundamental, which was reflected in the SC measurements. Roughly speaking, if the tongue position was more front (R_2 higher), the SC increased more steeply (at a greater rate) during R_1:f_o tuning; if the tongue position was more back (R_2 lower), the SC still increased during R_1:f_o tuning, but less steeply. In the current set of results, the SC started increasing unidirectionally around 600 Hz to 700 Hz; around D♯5 (nominally 622 Hz) to F♯5 (nominally 740 Hz). This pitch range roughly corresponds to sopranos' second *passaggio*, which is around E5 (nominally 659 Hz) to F♯5 (nominally 740 Hz) for sopranos (R. Miller, 1996). The results suggest that by the second *passaggio*, R_1:f_o

tuning occurs. The SC difference between [a] and [ɔ] was minimal in the upper range. This could be explained by the observation that both [a] and [ɔ] generally have relatively lesser distance between their R_1 and R_2. In higher pitches, their height (R_1) difference was potentially minimized by R_1:f_o tuning. As such, their SC differences remained minimal too. [ɛ] still showed a SC difference, which suggests that participants still attempted to differentiate the vowels as much as possible.

These results are comparable to those of Joliveau et al. (2004), where the average R_2 of [ɜ] (as in 'heard') appeared to be higher than the R_2 of the back vowels [ɑ], [ɔ] and [u], even in the highest pitch range.[6] As to the individual vowel differences, the differences observed in [ɛ] compared to [a] and [ɔ] is worth discussing. In the case of participants 2, 3 and 5, [ɛ] showed a higher predicted SC, revealing its frontness. However, [ɛ] had a systematically lower SC compared to [a] and [ɔ] for participants 1 and 6, suggesting that the vowels differed in height. The reduced SC for [ɛ] may be a result of [ɛ] having been systematically sung as a higher vowel compared to [a] and [ɔ]. Even if the R_2 of [ɛ] was high, if [ɛ] had a lower R_1 compared to [a] and [ɔ], the overall SC of [ɛ] may still be slightly lower. This interaction effect may also be why participant 4's estimated SC values of all three vowels were very similar. The results suggest that systematic vowel height differences between vowels may still occur at high fundamental frequencies. Given R_1:f_o tuning, the R_1 of [ɛ] could still be slightly lower. This is a predictable result, given that variations in vowel height have been found in Joliveau et al (2004) as well.

[6] A crucial difference between Joliveau et al. (2004) and the present one, however, is that the singers in Joliveau et al. (2004) sang a scale on one vowel, and repeated the scale on another vowel. On the other hand, participants sang all three vowels at a given pitch before repeating the three vowels in another pitch in the present study. This means that singers from Joliveau et al. (2004) were more likely to focus on timbre, projection and control, whereas the singers from the present study were more likely to focus on vowel quality and vowel contrasts, presumably due to the different natures of the tasks. Nonetheless, singers from both studies used R_1:f_o tuning, albeit with some differences between front and back vowels even at the highest pitches.

The behavior of SC before clear indication of R_1:f_o tuning could also be inferred with the A_1-A_2 analysis. R_1:f_o tuning would cause the f_o to become stronger, and the A_1-A_2 value would raise, while the SC would be close to f_o. During R_2 and $2f_o$ crossover, the A_1-A_2 would dip and raise a bit, while the SC may rise to a maximum during R_2 and $2f_o$ alignment and then fall. Based on Joliveau et al. (2004), R_1:f_o tuning tends to start around 500 Hz to 700 Hz depending on the vowel, and R_2 and $2f_o$ crossover may be roughly from 400 Hz to 800 Hz. Since the pitch range at which R_1:f_o tuning and R_2 and $2f_o$ crossover occurs overlap, SC is affected by both events. Nonetheless, rough correspondences were found between the SC and A_1-A_2 analysis. For example, the vowel [a] from participant 1 had an A_1-A_2 dip right before 600 Hz, which corresponded to an SC maximum as shown in figures 2 and 5. The vowel [ɔ] also had an A_1-A_2 dip at around 500 Hz, which corresponded roughly to a SC peak around 500 Hz as well. These values suggest the location of R_2 and $2f_o$ crossover, which corresponded to the results in Joliveau et al. (2004) for the same vowels.

The vowel modification strategy was not uniform across vowels. The frontness of [ɛ] tended to change before [a] and [ɔ] across all participants. A possible reason for this is that the inherent frontness of the vowel may affect the magnitude of its modification. While back vowels may have limited space for R_2 variation, front vowels may have a greater potential for variation and modification on the R_2 dimension. As [ɔ] lies in the posterior part of our mouths, there is inherently less space for it to vary in vowel backness. There is no clear consensus as to whether [a] in Cantonese is a front or back [a]. Nonetheless, the small range of space for [a] as a low vowel and its lower R_2 value implies that it cannot vary much in frontness. The present study shows that the vowel quality affects the vowel modification process, with the front vowel [ɛ] modifying before vowels [a] and [ɔ]. Further research may discern whether this pattern is widespread and why certain vowels modify earlier than others.

It is also worth comparing the spoken F_1 with the onset of the SC slope change to infer the vowel modification process before the second *passaggio*. The F_1 value for [ɛ] and [ɔ] in speech tended to be lower than the start of the SC slope change, which suggests that the R_1 of the sung tokens might have risen before the second *passaggio*. On the other hand, the "back" vowel [a] had a spoken F_1 value that is near or beyond a 1000 Hz. Theoretically speaking, [a] would not have needed to change their R_1 value at all, since it would not coincide with f_o. However, the vowel [a] still behaved like [ɛ] and [ɔ] in that the SC stayed close to the f_o around and beyond the second *passaggio*. This suggests that during the vowel modification process, the R_1 has been lowered. This finding partially supports Miller's (2000) description of the vowel modification process. He noted that when approaching the upper *passaggio* [ɛ] modifies towards [ɑ]; [ɑ] towards [ɔ]; and [ɔ] towards [o]. Miller's proposal suggests that vowels modify systematically: front vowels like [ɛ] raise their R_1, and back vowels like [ɑ] and [ɔ] lower their R_1. Our results also suggest that vowels modify systematically, and behavior is determined by height. The mid-vowels [ɛ] and [ɔ] raised its R_1, and the low vowel [a] lowered its R_1 prior to or around the second *passaggio*. This suggests that the vowel modification behavior is systematic and vowel quality dependent, and future acoustic or articulatory research considering a wider range of vowels may outline the exact process of such modification.

To summarize, every vowel was modified differently and the vowels' modification patterns correspond to Miller (2000) in that before the second *passaggio*, vowel modification was conditioned by height ([ɛ] and [ɔ] behaves similarly); after the second passaggio, vowel modification was conditioned by backness, i.e. the distance between R_1 and R_2 ([ɔ] and [a] behaved similarly). Although there is strong evidence of R_1:f_o tuning, vowels may still be slightly different in height. A perception task was conducted to test whether the intelligibility of the vowels reflected the results of the acoustic analysis.

Experiment Two

In the second experiment, native speakers of Cantonese listened to sound tokens sung by participant 1 from the first experiment. Participants listened to stimuli in the form of a discrimination task and identified the vowels they heard.

Methods and Materials

Participants

Seventy-nine self-reported Cantonese speakers with no formal vocal training in classical voice participated. The age range was between 18 and 50. Before the experiment, participants answered questions regarding their background in music, such as how often they listened to Western art vocal music, any choral experiences they had, and whether they have had other musical training including singing in popular genres. All participants signed an online consent form before the test. After consenting to participate, all participants first completed a screening task where they listened to speech stimuli from participant 1. Participants who did not perform well in discerning spoken vowels were not eligible to complete the rest of the experiment. This ensured all our participants could differentiate the singer's spoken vowels during speech, and that vowel misidentifications at higher pitches were therefore due to the change in fundamental frequency and vowel modification, and not due to any deficiency on the part of the listeners. Further details are outlined in the 'procedures' section below. The experiment protocol was approved by the Human Research Ethics Committee at <the authors' institute>.

Stimuli

Due to a concern about the length of the experiment, only a range of an octave from one participant was included in the perception test. Participant 1's recordings from the singing task were chosen as the stimuli for the perception test, because her results had the greatest difference in SC values across the three vowels, as demonstrated in figure 4. The pitch range was chosen based on pairwise comparisons between vowels' SC values. The SC difference across all three

vowel pairs peaked before 400 Hz, and the 95% credible intervals of [a] and [ɔ] touched zero by 440 Hz (A4), suggesting that confusions may start arising after this pitch. After 440 Hz, the SC difference between the vowels varied until around 800 Hz, where the difference between the vowels appear to stabilize. Thus, the pitch range included in the perception experiment spanned from A4 (nominally 440 Hz) to G♯5 (nominally 831 Hz). Only a subset of the spoken tokens was presented to the participants for the screening task. All vowels were included, but phonological environments were limited to the two preceding plosives, [p] and [t], while excluding [l], to reduce the length of the study.

The stimuli were presented in pairs, resembling an AX discrimination task, but answer choices were presented as an identification task. Token pairs included the following combinations: (1) same pitches with different vowels; (2) same pitches with same vowels; (3) different pitches with different vowels; and (4) different pitches with same vowels. Combinations (3) and (4) were presented at a fixed whole tone interval. For example, if the stimulus is [pɛ]-[pɔ] and [pɛ] is sung at B4 (nominally 494 Hz), then [pɔ] could only be either B4 (nominally 494 Hz, same pitch), A4 (nominally 440 Hz, one whole tone lower), or C♯5 (nominally 554 Hz, one whole tone above). The onset consonant was always kept constant. So that, for example, [pɛ]-[pɔ] was a possible stimulus, but [pɛ]-[tɔ] was not, because the onset consonant is not the same. Of the twelve frequencies, the whole tones from A4 (nominally 440 Hz, 494 Hz, 554 Hz, 622 Hz, 699 Hz, and 784 Hz) were used in combinations (3) and (4). The remaining tones (nominally 523 Hz, 587 Hz, 659 Hz, 740 Hz, and 831 Hz) were used in combinations (1) and (2). All vowel permutations ([a]-[ɔ], [ɔ]-[a], [a]-[ɛ], [ɛ]-[a], [ɔ]-[ɛ], [ɛ]-[ɔ], [a]-[a], [ɛ]-[ɛ], and [ɔ]-[ɔ]) were included. The tokens were amplitude-normalized by scaling the peak to 0.99, and duration-normalized to 680 ms, resulting in a normalized signing speed of 88 bpm. Since the singers in experiment one sang at a tempo of 88 bpm, duration normalization had almost no effect on the outcoming fundamental frequency. The speech

tokens were also normalized to a peak of 0.99 and a duration of 320 ms. A 500 ms gap was left in between the two tokens in each recording.

Participants were divided randomly into two groups to reduce the length of the experiment. Both groups were exposed to all 99 vowel-pitch combinations (eleven pitches x nine vowel combinations) but had different combinations of phonological environments. One group had 50 combinations with [t] onsets, and 49 with [p] onsets, the opposite holds true for the second group. Stimuli were presented in a randomized order. There were 396 recordings for the main task in total, and each participant answered 99 questions. Nine screening questions were included using the spoken tokens, which included all nine vowel permutations; five questions with a [t] onset and four with a [p] onset. In total, there were 405 recordings and each participant answered 108 questions. Table 1 summarizes the stimuli used in experiment 2.

Onset	Group 1		Group 2		
[p]	50		49		The stimulus order was either AB or BA per question
	AB	BA	AB	BA	
[t]	49		50		
	AB	BA	AB	BA	
Total	99 (198)		99 (198)		396 (+9 speech)

Table 1. Summary of the stimuli used in the perception experiment (Experiment 2)

Procedure

Participants were instructed to complete the experiment in a quiet environment via an online survey software, Qualtrics (2005). They were first asked their age and musical background and completed the screening task. They were instructed to play the stimuli on the screen at their own convenience and were presented with three answer choices in traditional Chinese characters per sound token in the recording. For every question, participants listened to two sung vowels. Therefore, participants were given two sets of three answer choices (6 identification answer choices per question). The answer choices had no mismatches with consonant onsets. For example, if the stimulus was [pa] and [pa], the answer choices presented

were “巴” ([pa:˥]), “啤” ([pɛ:˥]), and “波” ([pɔ:˥]) for both tokens. Only participants who answered seven out of the nine screening questions correctly moved on to the main task, otherwise, they were excluded. Of the 79 participants, 4 were excluded due to their failure in the screening task. The procedure of the main task was same as the screening task. If participants wished to listen to the same recording more than once, they were free to replay the stimuli. All questions were presented in a randomized order.

Stimulus presentation resembled an AX discrimination task, but participants were asked to identify the pairs of vowels like an identification task. This allowed us to gather both data on the intelligibility of the vowels themselves, and data on inferred discriminability between vowels. As an example of interpretation of the discriminability test, if the tokens presented were [pa] and [pɔ], and the participant discriminated them, we infer that they could discriminate [pa] and [pɔ], but it could be that they misidentified the tokens as [pɛ] and [pa] respectively. As identification and discrimination analyses tasks may give different results in the way described, our analyses included both real identification and inferred discrimination results.

Results

The mean identification rates are summarized in Figure 6. Error bars represent 95% Wald confidence intervals[7] (Devleesschauwer et al., 2015). As shown, only the results for [ɛ] support our hypothesis: the identification rate of [a] and [ɔ] fluctuated greatly across the entire pitch range, while the identification rate of [ɛ] started dropping at E5 (nominally 659 Hz), and dropped drastically after F♯5 (nominally 740 Hz) onwards. Fixed effects of vowel and fundamental frequency on the effect of correct identification of vowels were evaluated using a logistic regression model in R (Bates et al., 2015; R Core Team, 2020). For the identification

[7] Note that the error bars in Figures 2 and 4 are of credible intervals (Bayesian), which should not be directly compared with confidence intervals (frequentist) shown in the results of experiment two (Figures 6 and 7).

model, two-way interactions between vowel (three levels) and fundamental frequency (twelve numeric levels) were included. The reference level was the vowel [a]. Results show that the identification rates of [ɔ] were significantly different from the reference level ($\beta = 0.564$, SE = 0.253, $z = 2.235$, $Pr(>|z|) = 0.025$). [ɔ]-[ɛ]. The identification rates of [ɛ] were also significantly different from the reference level ($\beta = 15.740$, SE = 0.740, $z = 21.272$, $Pr(>|z|) < 0.001$). There was no main effect of fundamental frequency ($\beta = -0.0001$, SE = 0.0002, $z = 0.531$, $Pr(>|z|) = 0.596$). However, there were significant interaction effects between fundamental frequency and vowel [ɔ] ($\beta = -0.002$, SE = 0.0004, $z = -4.270$, $Pr(>|z|) < 0.001$) and vowel [ɛ] ($\beta = -0.020$, SE = 0.0009, $z = -20.793$, $Pr(>|z|) < 0.001$) respectively.

[insert Figure 6.]

FIG. 6. Percentage of correct identification responses. x-axis: fundamental frequency; y-axis: percentage of correct identification responses; error-bars: 95% Wald confidence intervals.

Figure 7 shows the mean inferred discrimination results, with error bars representing 95% Wald confidence intervals. As Figure 7 shows, the inferred discrimination results largely corresponded to the identification results in that the pairs including the vowel [ɛ] were distinguished well: [a]-[ɛ] and [ɔ]-[ɛ] pairs were distinguished well up to F5 (nominally 699 Hz), and inferred discrimination rates started dropping drastically after G5 (nominally 784 Hz). Apart from 659 Hz, 699 Hz, and 784 Hz, the confidence intervals of the participants' mean responses for [a]-[ɔ] crosses 50%, suggesting minimal discrimination between the two vowels. Fixed effects of vowel combination and fundamental frequency on the effect of successful inferred discrimination of different vowels were evaluated using a logistic regression model in R (Bates et al., 2015; R Core Team, 2020). Two-way interactions between vowel combinations (three levels) and fundamental frequency (eleven numeric levels) were included. The vowel combination [a]-[ɔ] was the reference level. Results show that the vowel combinations [a]-[ɛ]

(β = 26.394, SE = 2.917, z = 9.048, Pr(>|z|) < 0.001) and [ɔ]-[ɛ] (β = 20.367, SE = 2.228, z = 9.140, Pr(>|z|) < 0.001) were significantly different from the reference level. There was a main effect of fundamental frequency (β = 0.001, SE = 0.0004, z = 2.292, Pr(>|z|) < 0.022). There were also significant interaction effects between fundamental frequency and vowel combinations [a]-[ɛ] (β = -0.032, SE = 0.004, z = -8.749, Pr(>|z|) < 0.001) and [ɔ]-[ɛ] (β = -0.024, SE = 0.003, z = -8.556, Pr(>|z|) < 0.001) respectively.

[insert Figure 7.]

FIG. 7. Percentage of correct inferred discrimination responses between vowel pairs. x-axis: fundamental frequency; y-axis: percentage of correct inferred discrimination responses; error-bars: 95% Wald confidence intervals.

Discussion

In the perception task, all three vowels, [a], [ɔ], and [ɛ], showed identification rates above chance. But the identification rates of [ɔ] did not show a clear trend, except that the vowel was better identified at fundamental frequencies lower than 494 Hz, and less identifiable at fundamental frequencies higher than 740 Hz. The identification rates of [a] consistently stayed above chance, with peak identification rates being around 587 Hz to 622 Hz. This frequency range corresponded roughly to the differences in modelled SC estimates (Figure 4), where the peak difference between the two vowels also lay between 587-622 Hz. The observed pattern seemed to suggest that when [a] and [ɔ]'s SC differences were higher, [a] identification rates increased, but [ɔ] was more ambiguous to listeners. Interestingly, their inferred discrimination rates were not affected by this peak in SC difference. The identification rate of [ɛ] was closest to our predictions. Identification rates were close to 100% up to 659 Hz, where rates of correct identification responses started to slowly drop. Vowel pairs that involve [ɛ] (i.e. [a]-[ɛ] and [ɔ]-[ɛ]) roughly corresponded to the predicted SC differences as well. In figure 4, the credible intervals of [a]-[ɛ] and [ɔ]-[ɛ] touched 0 at 784 Hz and 740 Hz respectively. In figure 7, the

inferred discrimination rates for both pairs of vowels started to decrease at 659 Hz, and decreased drastically after 784 Hz. Results show that SC measurements do roughly correspond with vowel intelligibility results.

While it was hypothesized that there would be both a gradual change in resonances and perceptibility depending on f_o rise, the results of the perception test show that the gradual change in perceptibility was not an overall tendency, and was specific to certain phones, namely [ɛ] in the current study. This tells us that we cannot make a single generalization for the behavior of all the vowels when it comes to singing. One way to interpret why only [ɛ] was gradually affected by an f_o rise is that, in speech, [a] and [ɔ] both have R_1 and R_2 close together. Even when pitch is not very high, listeners are therefore likely to interpret similar harmonics (e.g. f_o and $2f_o$) as if they are R_1 and R_2 for both [a] and [ɔ]. Furthermore, height information (R_1) becomes unavailable due to both the increasing pitch and R_1:f_o tuning. The consequent lack of unambiguous formant and resonant information could increase confusability for [a] and [ɔ]. On the other hand, [ɛ] generally has a higher R_2, and is likely to have maintained its high R_2 in the midrange of the fundamental frequency before it slowly dropped, which was reflected by the A_1-A_2 analysis. If so, listeners might have attended to higher harmonics in the mid-range for [ɛ] for a frontness distinction.

Overall Discussion

Singing differs from other forms of music-making; singers convey lyrics. While both the acoustics of singing and speech have each been explored on their own terms, how the unique acoustics of singing affect the acoustics of basic speech sounds has yet to be fully understood. This study has explored the acoustic effects of vowel modification. This is a topic of interest since vowel modification is necessary for singers to maintain voice projection and timbral control as the fundamental frequency increases, even if vowel modification translates to decreasing vowel intelligibility. The existing literature on resonance tuning in sopranos and

pedagogical research on vowel modification has led us to expect R_1:f_o tuning. Accordingly, we asked how the vowels change with f_o raising, whether SC measurements could capture this process acoustically, and how vowel intelligibility is compromised. The results confirmed the expected resonance tuning strategy in sopranos with R_1:f_o tuning at higher registers and revealed several implications as outlined below.

First, vowel modification is not a unified process. Different vowels behave differently during vowel modification, and its subsequent effects on vowel intelligibility therefore also differ. While some existing studies on resonance tuning have included multiple vowels (Henrich et al., 2011; Joliveau et al., 2004), their focus has not been on the nature of vowel modification, nor on the subsequent effects on intelligibility. In the current set of results, we have found that [ɛ] tended to start modifying at comparatively lower pitches, yet its intelligibility rates tended to stay high even at higher pitches. On the contrary, [a] and [ɔ] started modifying only at higher pitches, yet their confusability tended to be high in lower registers too. Recent studies have proposed that in non-operatic styles, the so-called 'point vowels' /i/, /u/ and /a/ may act as acoustic landmarks while other vowels demonstrate decreasing intelligibility (Friedrichs et al., 2017). The present study does not demonstrate the same tendencies, as [a]-[ɔ] confusions happened early in the pitch range. We attribute these results to vowel modification being conditioned by vowel frontness when the upper register is concerned. Since front vowels have a greater range of articulatory space available, front vowels may start modifying earlier while keeping their perceptual differences with other vowels for longer. Further research will be required to conclude whether vowel modification patterns are bounded by features and/or determined by vocal registers.

Second, vowel modification is not motivated by f_o-F_1 adjacency. While one might expect that R_1 must be modified because the f_o passes through ranges like that of F_1 in speech, the current set of results demonstrates otherwise. There is a considerable distance between the

mean F_1 in speech and the onset of SC-to-f_o alignment. This suggests that the vowels' R_1 underwent modification in lower pitches before SC-to-f_o alignment. In other words, vowel height did not remain static until the f_o was as high as R_1. The case of [a] is especially apparent, since its spoken mean F_1 is typically very high, providing evidence that R_1 modification is not motivated by f_o-F_1 adjacency. We may speculate that vowel modification is motivated either by R_1:f_o tuning, or by external reasons for modifications in lower registers. These findings echo existing research that show R_1 adjustments prior to R_1:f_o tuning (Joliveau et al., 2004). Further research may consider the motivations to vowel modification in different registers, as well as the role of R_2 modification.

Third, our current methodology relied on minimized vowel height differences during R_1:f_o tuning to interpret SC results. Results show that vowel height differences may still be reflected in the acoustic signal during R_1:f_o tuning. In the production experiment, participants were asked to sing all three vowels in the same breath group, which should demotivate them from changing their jaw height across vowels at every given pitch. Therefore, we did not expect the R_1 of each vowel to differ much. Nevertheless, two participants show lower SC values for [ɛ] compared to [a] and [ɔ] during R_1:f_o tuning, which appears to be systematic. This calls into question whether there are systematic differences in R_1:f_o tuning strategies for different vowels.

This study has investigated how sopranos modify their vowels when approaching higher pitches, and how vowel intelligibility is affected by the vowel modification process. All six sopranos demonstrated fluctuating SCs in the lower and middle register, reflecting fluid R_1 and R_2 change in the process. As the upper register approaches, R_1:f_o tuning is evident in all three vowels, though the singers still attempt to differentiate the vowels. Comparisons with resonance measurements in speech suggest that vowel modification started to occur before SC-to-f_o alignment, and its patterns are likely to support what existing singers have theorized. Perception results showed high confusability between the two back vowels, with decreasing

intelligibility of [ɛ], at a pitch that roughly corresponded to decreasing SC differences between vowels. The results suggest that future research into resonance and timbre in soprano voices should take a more nuanced approach to possible differential tuning strategies for different vowels.

Acknowledgements

References

Andersen, N. (1974). On the calculation of filter coefficients for maximum entropy spectral analysis. *GEOPHYSICS*, *39*(1), 69–72. https://doi.org/10.1190/1.1440413

Assmann, P. F. (1991). The perception of back vowels: Centre of gravity hypothesis. *The Quarterly Journal of Experimental Psychology Section A*, *43*(3), 423–448. https://doi.org/10.1080/14640749108400980

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

Bauer, R. S., & Benedict, P. K. (1997). *Modern Cantonese Phonology*. Walter de Gruyter.

Beddor, P. S., & Hawkins, S. (1990). The influence of spectral prominence on perceived vowel quality. *The Journal of the Acoustical Society of America*, *87*(6), 2684–2704. https://doi.org/10.1121/1.399060

Benolken, M. S., & Swanson, C. E. (1990). The effect of pitch-related changes on the perception of sung vowels. *The Journal of the Acoustical Society of America*, *87*(4), 1781–1785. https://doi.org/10.1121/1.399426

Boersma, P., & Weenink, D. (2018). *Praat: Doing phonetics by computer* (6.0.43) [Computer software]. http://www.praat.org/

Bozeman, K. W. (2013). *Practical Vocal Acoustics: Pedagogic Applications for Teachers and Singers*. Pendragon Press.

Bunch, M., & Chapman, J. (2000). Taxonomy of singers used as subjects in scientific research. *Journal of Voice*, *14*(3), 363–369. https://doi.org/10.1016/S0892-1997(00)80081-8

Callaghan, J. (2000). *Singing and voice science*. Singular Pub. Group.

Chistovich, L. A., & Lublinskaya, V. V. (1979). The 'center of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. *Hearing Research*, *1*(3), 185–195. https://doi.org/10.1016/0378-5955(79)90012-1

Delattre, P., Liberman, A. M., Cooper, F. S., & Gerstman, L. J. (1952). An Experimental Study of the Acoustic Determinants of Vowel Color; Observations on One- and Two-Formant Vowels Synthesized from Spectrographic Patterns. *Word*, *8*(3), 195–210. https://doi.org/10.1080/00437956.1952.11659431

Deme, A. (2017). The Identification of High-pitched Sung Vowels in Sense and Nonsense Words by Professional Singers and Untrained Listeners. *Journal of Voice*, *31*(2), 252.e1-252.e14. https://doi.org/10.1016/j.jvoice.2016.07.008

Deng, L., & O'Shaughnessy, D. (2003). *Speech Processing: A Dynamic and Optimization-Oriented Approach*. CRC Press.

Devleesschauwer, B., Torgerson, P., Charlier, J., Levecke, B., Praet, N., Roelandt, S., Smit, S., Dorny, P., Berkvens, D., & Speybroeck, N. (2015). *prevalence: Tools for prevalence assessment studies.* [R package version 0.4.0]. http://cran.r-project.org/package=prevalence

Doscher, B. M. (1994). *The functional unity of the singing voice* (2nd ed). Scarecrow Press.

Fox, R. A., Jacewicz, E., & Chang, C.-Y. (2011). Auditory Spectral Integration in the Perception of Static Vowels. *Journal of Speech, Language, and Hearing Research*, *54*(6), 1667–1681. https://doi.org/10.1044/1092-4388(2011/09-0279)

Friedrichs, D., Maurer, D., & Dellwo, V. (2015). The phonological function of vowels is maintained at fundamental frequencies up to 880 Hz. *The Journal of the Acoustical Society of America*, *138*(1), EL36–EL42. https://doi.org/10.1121/1.4922534

Friedrichs, D., Maurer, D., Rosen, S., & Dellwo, V. (2017). Vowel recognition at fundamental frequencies up to 1 kHz reveals point vowels as acoustic landmarks. *The Journal of the Acoustical Society of America*, *142*(2), 1025–1033. https://doi.org/10.1121/1.4998706

Gottfried, T. L., & Chew, S. L. (1986). Intelligibility of vowels sung by a countertenor. *The Journal of the Acoustical Society of America*, *79*(1), 124–130. https://doi.org/10.1121/1.393635

Hashimoto, O. Y. (1972). *Phonology of Cantonese*. Cambridge University Press.

Hayward, K. (2016). *Experimental Phonetics: An Introduction*. Routledge.

Henrich, N., Smith, J., & Wolfe, J. (2011). Vocal tract resonances in singing: Strategies used by sopranos, altos, tenors, and baritones. *The Journal of the Acoustical Society of America*, *129*(2), 1024–1035. https://doi.org/10.1121/1.3518766

Joliveau, E., Smith, J., & Wolfe, J. (2004). Vocal tract resonances in singing: The soprano voice. *The Journal of the Acoustical Society of America*, *116*(4), 2434–2439. https://doi.org/10.1121/1.1791717

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation.

2008 IEEE International Conference on Acoustics, Speech and Signal Processing, 3933–3936. https://doi.org/10.1109/ICASSP.2008.4518514

Kris, T., & S. Greg, T. (1997). Spectral Properties of Fricatives in Amyotrophic Lateral Sclerosis. *Journal of Speech, Language, and Hearing Research*, *40*(6), 1358–1372. https://doi.org/10.1044/jslhr.4006.1358

Lewcock, R., Pirn, R., Meyer, J., Hutchins, C. M., Woodhouse, J., Schelleng, J. C., Richardson, B., Martin, D. W., Benade, A. H., Campbell, M., Rossing, T. D., & Sundberg, J. (2001). *Acoustics* (Vol. 1). Oxford University Press. https://doi.org/10.1093/gmo/9781561592630.article.00134

Matthews, S. J., & Yip, V. (1994). *Cantonese: A Comprehensive Grammar*. Routledge.

Miller, D. G. (2000). *Registers in Singing. Empirical and Systematic Studies in the Theory of the Singing Voice*. Ponsen & Looijen BV, Wageningen.

Miller, R. (1996). *The structure of singing: System and art in vocal technique*. Schirmer.

Miller, R. (2000). *Training soprano voices*. Oxford University Press.

Monsen, R. B., & Engebretson, A. M. (1983). The Accuracy of Formant Frequency Measurements: A Comparison of Spectrographic Analysis and Linear Prediction. *Journal of Speech, Language, and Hearing Research*, *26*(1), 89–97. https://doi.org/10.1044/jshr.2601.89

Nittrouer, S., Studdert-Kennedy, M., & McGowan, R. S. (1989). The Emergence of Phonetic Segments. *Journal of Speech, Language, and Hearing Research*, *32*(1), 120–132. https://doi.org/10.1044/jshr.3201.120

Peirce, J., & MacAskill, M. (2018, May). *Building Experiments in PsychoPy*. SAGE Publications Ltd.

Qualtrics (July 2019). (2005). [Computer software]. Qualtrics. https://www.qualtrics.com

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Shue, Y.-L. (2010). *The voice source in speech production: Data, analysis and models*. University of California, Los Angeles.

Sibelius First (2018.6). (2006). [Computer software]. Avid Technology Europe Limited. https://my.avid.com/get/sibelius-first

Smith, L. A., & Scott, B. L. (1980). Increasing the intelligibility of sung vowels. *The Journal of the Acoustical Society of America*, *67*(5), 1795–1797. https://doi.org/10.1121/1.384308

Sundberg, J. (1974). Articulatory interpretation of the "singing formant." *The Journal of the Acoustical Society of America*, *55*(4), 838–844. https://doi.org/10.1121/1.1914609

Sundberg, J. (1975). Formant Technique in a Professional Female Singer. *Acta Acustica United with Acustica*, *32*(2), 89–96.

Sundberg, J., Lã, F. M. B., & Gill, B. P. (2013). Formant Tuning Strategies in Professional Male Opera Singers. *Journal of Voice*, *27*(3), 278–288. https://doi.org/10.1016/j.jvoice.2012.12.002

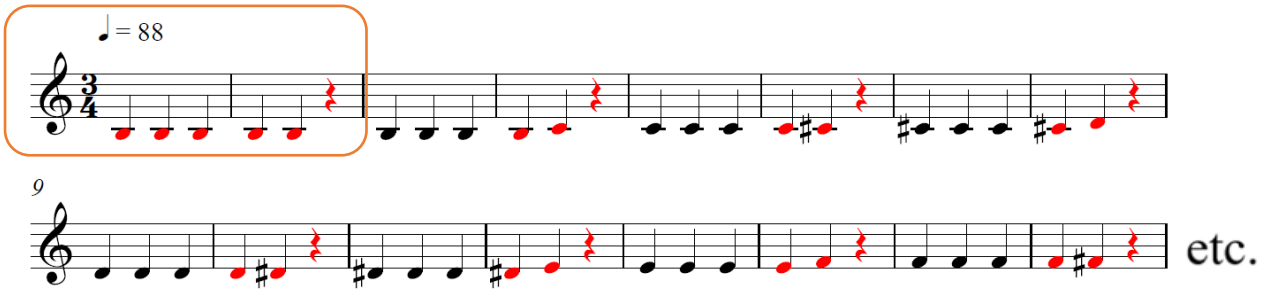
Titze, I. R., Baken, R. J., Bozeman, K. W., Granqvist, S., Henrich, N., Herbst, C. T., Howard, D. M., Hunter, E. J., Kaelin, D., Kent, R. D., Kreiman, J., Kob, M., Löfqvist, A., McCoy, S., Miller, D. G., Noé, H., Scherer, R. C., Smith, J. R., Story, B. H., … Wolfe, J. (2015). Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization. *The Journal of the Acoustical Society of America*, *137*(5), 3005–3007. https://doi.org/10.1121/1.4919349

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279

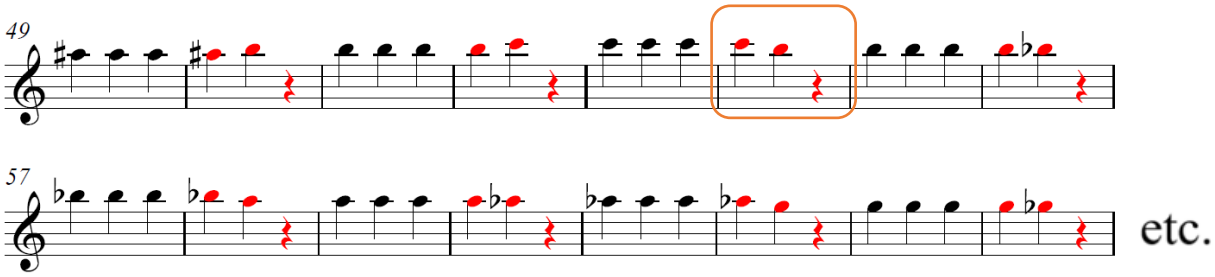
Wood, S. N. (2019). *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation* (1.8-31) [Computer software]. https://CRAN.R-project.org/package=mgcv

Instructions:

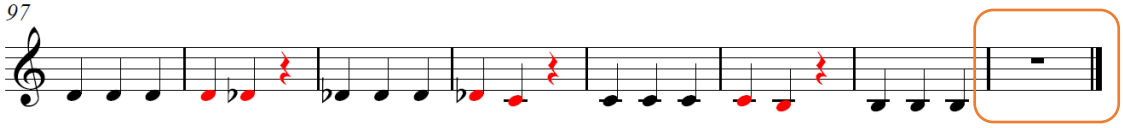
- You will be singing the following warm up exercise using a series of different syllables.
- An audio file will be played in the background to give you the guiding pitch and to keep the tempo.
- Please do **not** sing the notes in RED, which will still be played in the audio.
- Please sing the notes in BLACK, with the specific syllable.
- Feel free to let me know if you would want to take a rest or drink water in between exercises!
- The first 2 bars do not need to be sung – the purpose of the first two bars is only to give you the guiding pitch and to set the tempo.



- After reaching High C, the background audio file will guide you to sing the descending scale downwards.



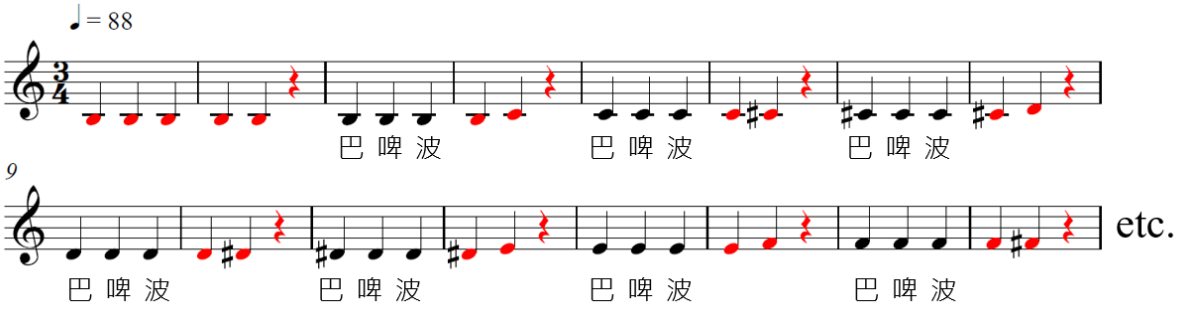
- When the audio file no longer gives you the following semitone downwards, and you hear a full bar rest. It indicates that the exercise is completed.



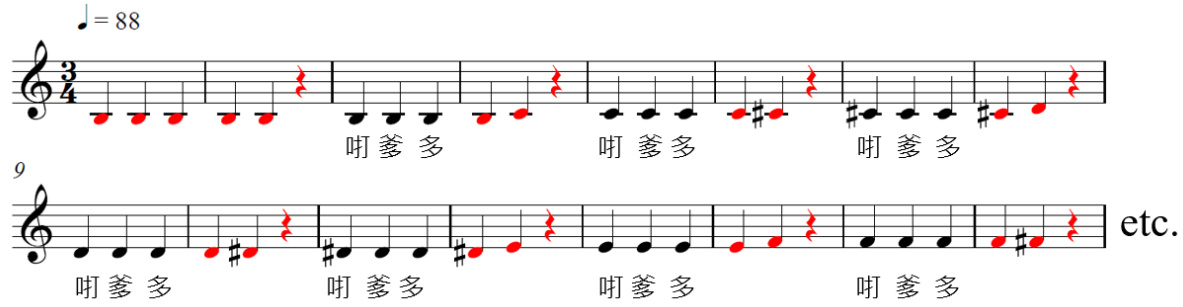
- Most of the text are written in Cantonese characters and are listed below with Cantonese Jyutping. Please let me know if there are any words you don't know.

	b	d	l
aa1	巴/爸 [baa1]	打/咑 (as in 一打) [daa1]	啦 [laa1]
e1	啤 [be1]	爹 [de1]	咧/哩 [le1]
o1	波 [bo1]	多 [do1]	囉 [lo1]

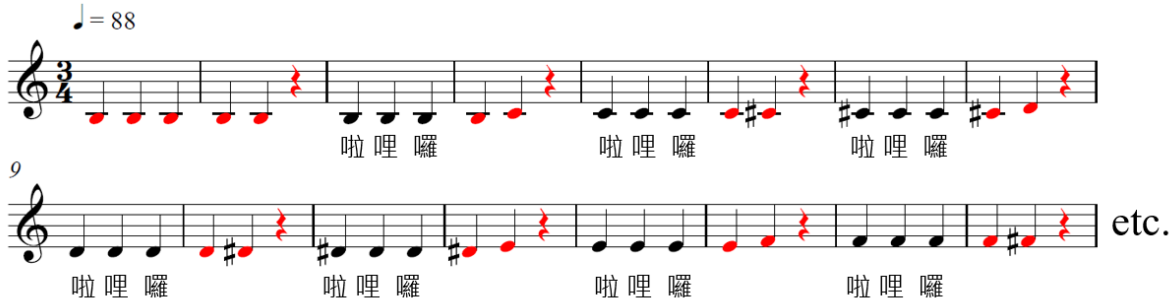
baa1 – be1 – bo1



daa1 – de1 – do1



laa1 – le1 – lo1



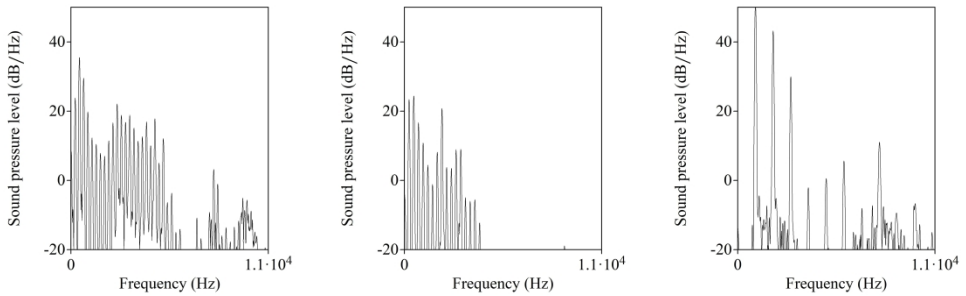


Figure 1. Narrowband spectra (window length = 0.03 s) of the vowel [ɛ] spoken at 233 Hz (left), sung at 262 Hz (middle), and sung at 988 Hz (right). Figures were generated in Praat (Boersma & Weenink, 2018).

2540x846mm (72 x 72 DPI)

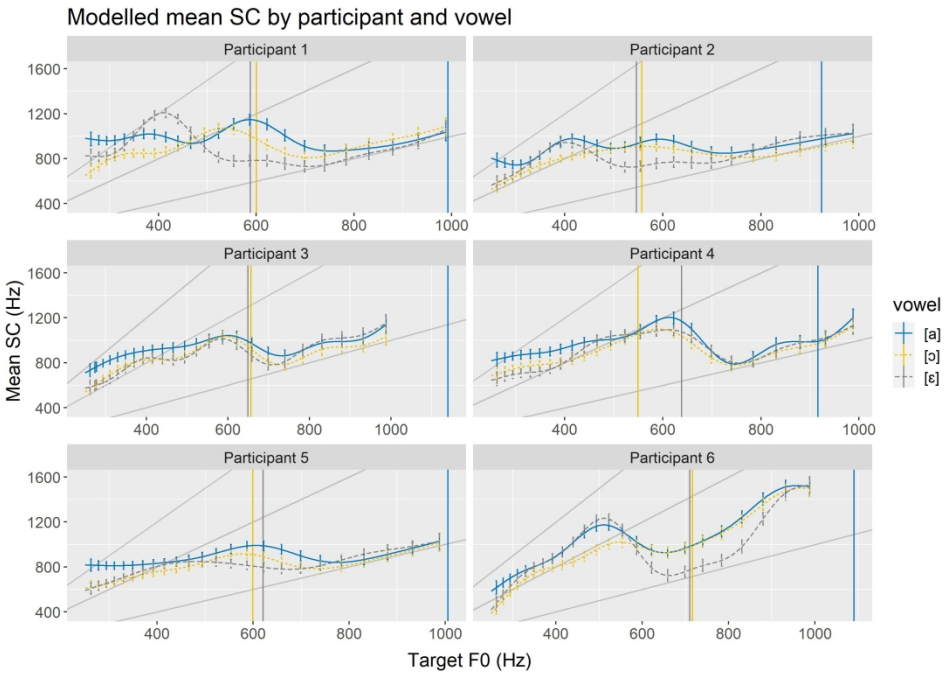


FIG. 2. Estimated mean SC by participant and vowel. x-axis: target fundamental frequency in Hz; y-axis: modelled mean spectral centroid in Hz; error bars: 95% credible intervals (calculated using the results of the GAMM). The diagonal grey lines show fo to 3fo. The participants' average speech F1s are included as vertical lines.

279x203mm (300 x 300 DPI)

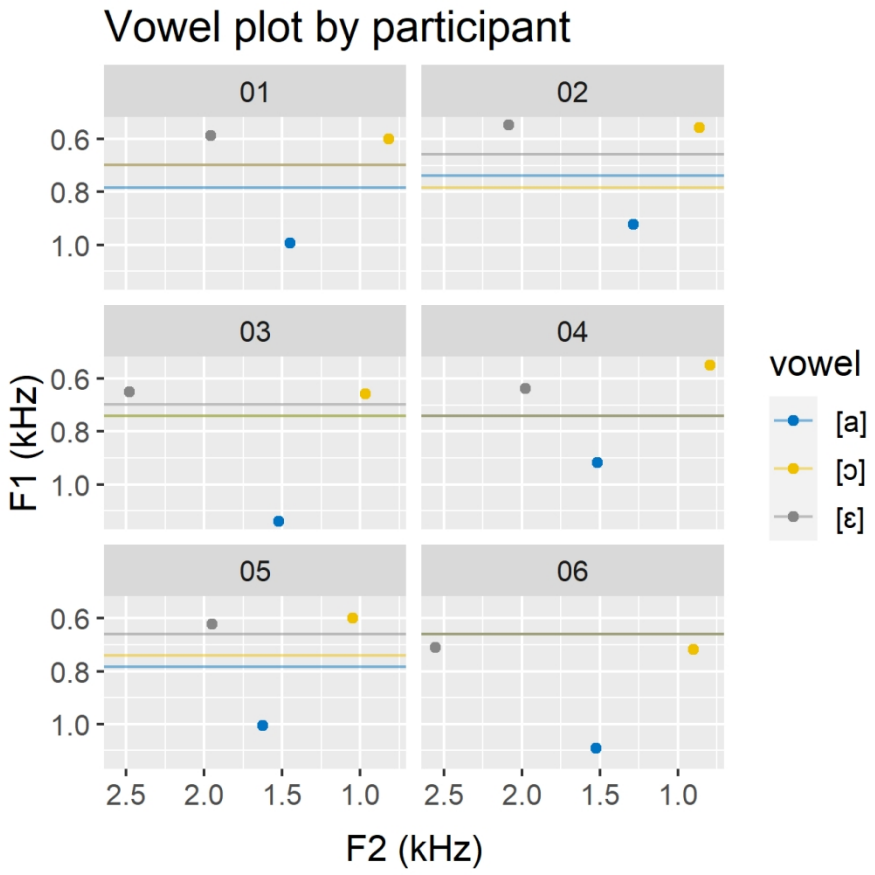


FIG. 3. F1 and F2 of the spoken vowels. x-axis: F2 in Hz; y-axis: F1 in Hz. y-intercept: minimum derivative of the SC between the 650-850 Hz range (~onset of R1:fo tuning)

322x322mm (118 x 118 DPI)

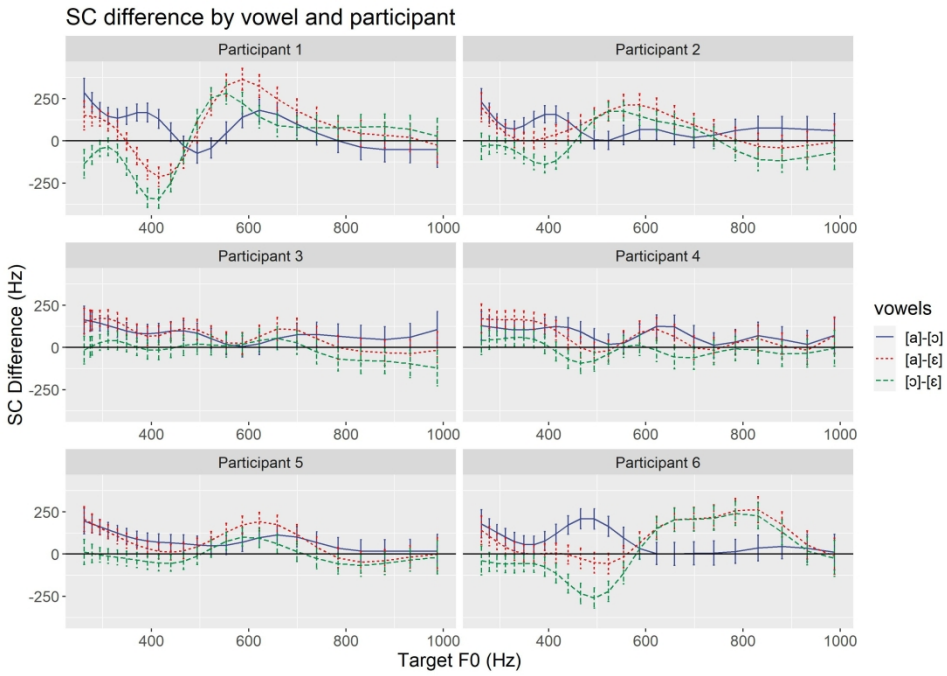


FIG. 4. Estimated SC differences in vowel pairs. x-axis: target fundamental frequency in Hz; y-axis: modelled spectral centroid difference in Hz; error bars: 95% credible intervals.

279x203mm (300 x 300 DPI)

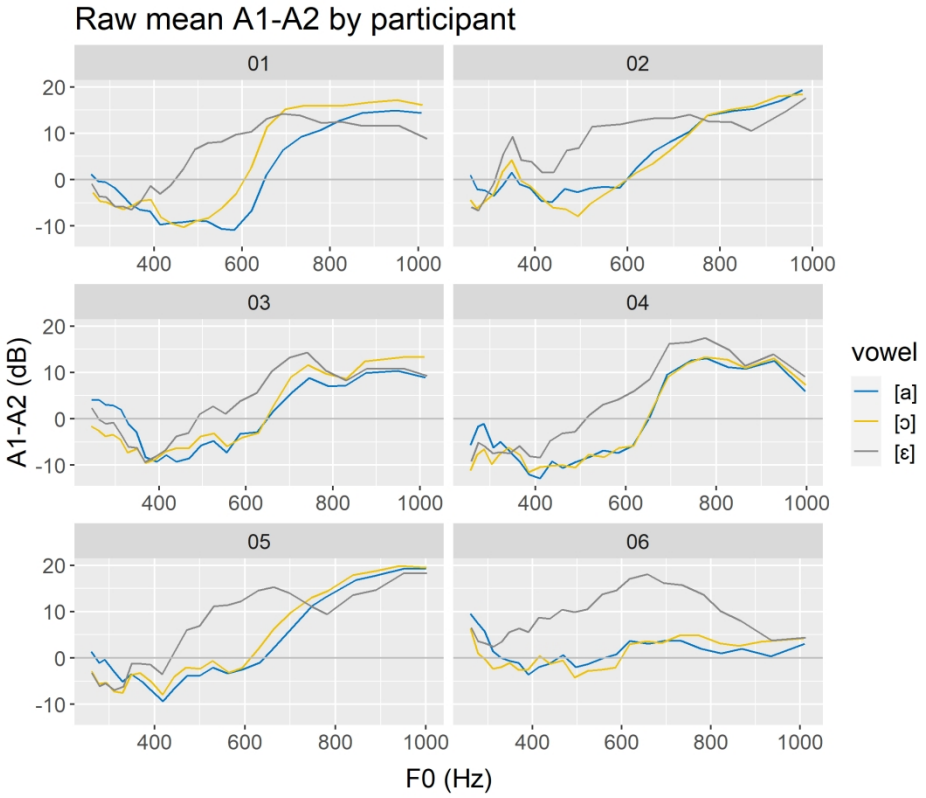


FIG. 5. A1-A2 by participant. x-axis: measured F0 in Hz; y-axis: difference in amplitude between the first and second harmonics.

516x452mm (118 x 118 DPI)

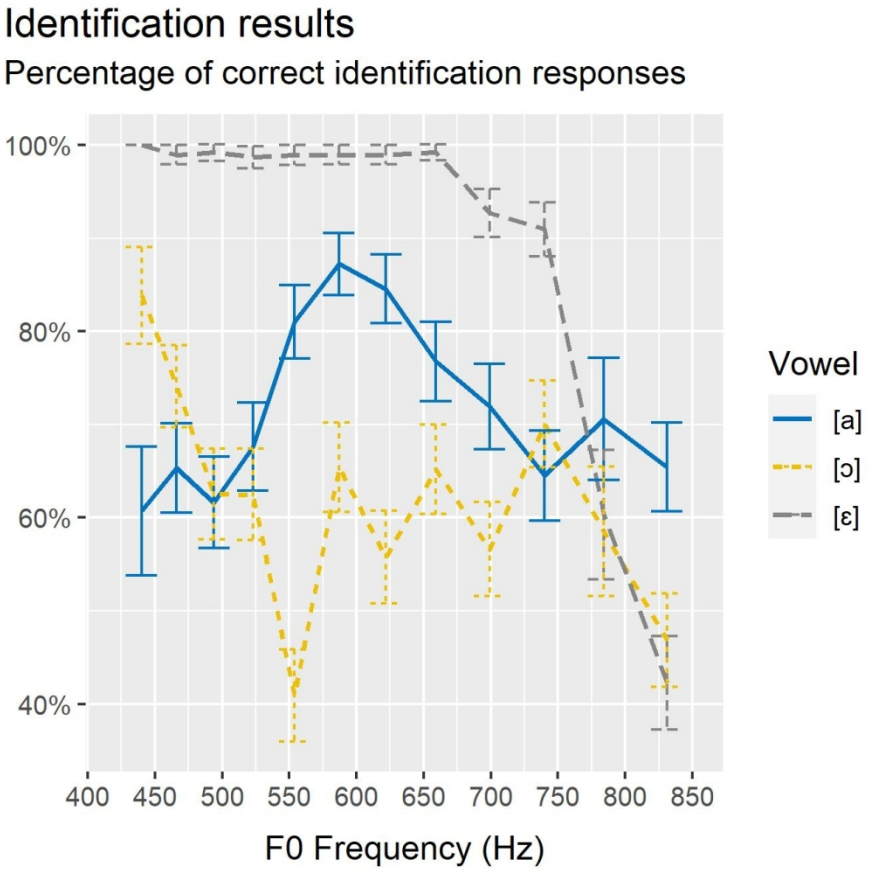


FIG. 6. Percentage of correct identification responses. x-axis: fundamental frequency; y-axis: percentage of correct identification responses; error-bars: 95% Wald confidence intervals.

127x127mm (300 x 300 DPI)

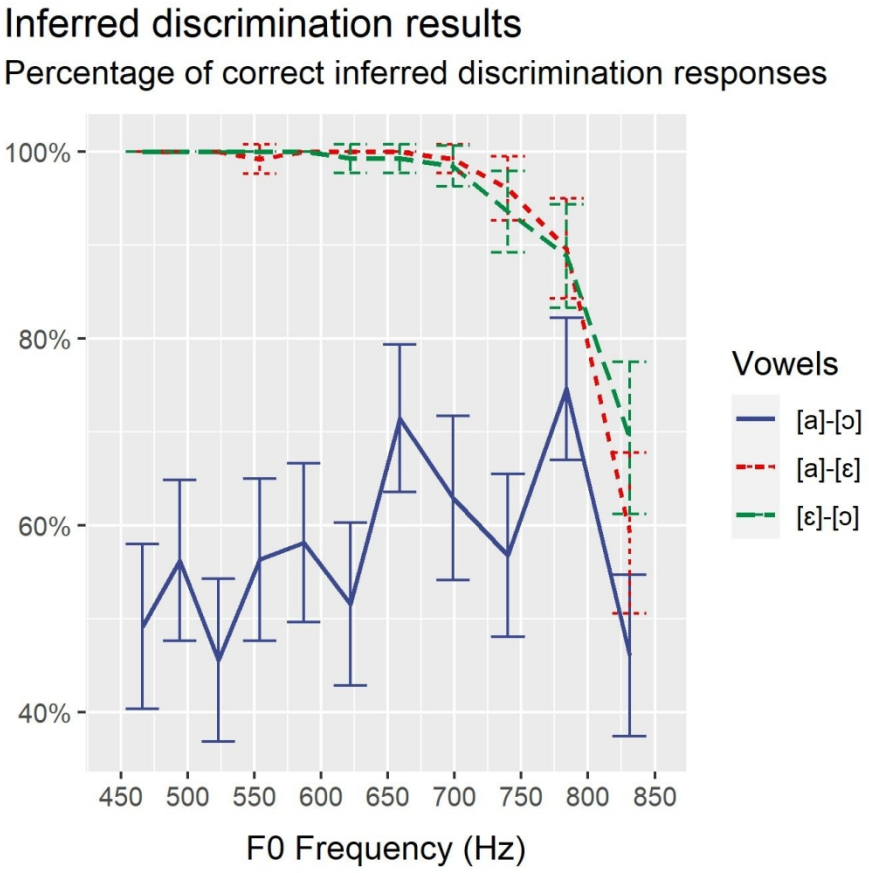


FIG. 7. Percentage of correct inferred discrimination responses between vowel pairs. x-axis: fundamental frequency; y-axis: percentage of correct inferred discrimination responses; error-bars: 95% Wald confidence intervals.

127x127mm (300 x 300 DPI)