Interspeech 2024
1-5 September 2024, Kos, Greece

Pitch-driven adjustments in tongue positions: Insights from ultrasound imaging

May Pik Yu Chan[1], *Jianjing Kuang*[1]

[1]Department of Linguistics, University of Pennsylvania, USA
`pikyu@sas.upenn.edu, kuangj@ling.upenn.edu`

Abstract

As an important singing technique, trained singers tend to adjust their resonance space based on pitch, especially for higher pitches. The mechanism behind vowel modification remains unclear - is it learned only for certain acoustic goals, or inherently required for articulatory ease? To address this question, we explore whether speakers adjust their resonance spaces when targeting different pitches; and if so, how. 23 participants participated in a production experiment involving ultrasound tongue imaging technique. Participants were asked to sing vowels across their pitch range rising by semitone. Results show that there is a link between pitch range and tongue positional adjustments - participants with a larger pitch range tended to adjust their tongue positions as they reached higher pitches, including some individuals who reported no history of vocal training. We discuss our results with respect to interactions between the source and filter in the human voice production system.

Index Terms: vowel-pitch interaction, source-filter theory, vowel modification, ultrasound tongue imaging

1. Introduction

Speech production is limited to a very narrow region at the lower end of ones' pitch range. While average speakers have at least an octave of a working pitch range [1], the general pitch range used in everyday speech is much smaller, ranging from around 90∼160 Hz and 160∼250 Hz for male and female speakers respectively [2]. Among various possible challenges that could arise from exploring ones' higher pitches, one is that navigating a intended and consistent tone quality across pitch ranges comes with great difficulty.

One group of speakers who are exceptional at navigating different regions of their pitch range is singers. An important source of the effectiveness of singers' voice production is their ability to adjust their resonance space depending on the pitch. This technique, "formant tuning" – also known as "vocal tract tuning" or "resonance tuning", describes a process in which vocal tract adjustments are made such that specific low resonances ("formants") are in close proximity to specific harmonics (often the fundamental), which enables a boost in energy in the desired frequency regions of the spectrum [3, 4, 5, 6, 7]. Resonance tuning has been found to be particularly important in the case of higher fundamental frequencies (F0) [8, 9, 6].

There are several reasons why singers engage in resonance tuning. In some cases, F0 could be sung at a frequency that is higher than a spoken vowel formant at regular vocal efforts (e.g., singing /i/ at a pitch higher than 300∼400 Hz, its F1 range), leaving the first formant with no harmonic to support. To overcome this challenge, sopranos are often found to avoid singing high vowels at a very high pitch where their first resonance could become lower than the fundamental, and in turn tune their first resonance close to the fundamental frequency instead [6, 7, 10]. This also provides a boost in energy of the F0 region to increase vocal loudness. Apart from sopranos, resonance tuning has also been reported in male voices [7, 5], where it occurs more often around the *passaggio*, to help navigate the region between the upper end of the "chest register" towards the "falsetto" in singing [11, 12]. Since the adjustment of vocal tract space interacts with vowel formant structure, the process of resonance tuning is sometimes understood by singers as "vowel modification", where vowel contrasts alter in the resonance space through a chain-shift process by pitch, often reaching vowel contrast neutralization in higher ranges [10, 13].

While the theories on why resonance tuning is effective for singers circle around the Western classical tradition [5, 6], the application of resonance tuning has been found across different musical traditions, ranging from classical, traditional and folk [14, 15, 16] to contemporary commercial music or musical theatre [17, 18, 19, 20, 21], where the main differences lie in which harmonic-resonance combinations are preferred. Together, this suggests that different stylistic traditions have (to some degree) independently developed the understanding that changes in the resonance structure could facilitate the production of different pitch targets given specific acoustic targets. With this background in mind, one might wonder whether there is an inherent motivation to vocal tract adjustments by pitch ranges beyond stylistic preferences determined by musical style.

A core question remains: what is the mechanism behind vowel modification? Do speakers only engage in vowel modification when they have given acoustic goals in singing (e.g. boosting the F0 with the first resonance)? Or is vowel modification inherently required for articulatory ease at higher pitch ranges? In this study, we explore whether non-professional singers who have lesser or no vocal training might also engage in vowel modification-like strategies when pursuing pitches outside of their normal speaking range. If it is only experienced singers who engage in vocal tract adjustments in singing, that suggests that vowel modification might be a technique inherent to certain acoustic goals pertaining to the singers' musical background. On the other hand, if non-singers are also found to engage in vocal tract adjustments when approaching higher pitches, then it suggests that the modification of the filter structure to support voice source characteristics might be inherently required for effective voice production engrained in our articulatory mechanism. This study therefore focuses on non-professional singers — if it is necessary to change the vowels to reach higher notes, then non-professional singers will also use comparable strategies as singers when singing higher pitches. This topic raises interesting questions on how and when voice

10.21437/Interspeech.2024-2247

users may choose to interact information between the voice source and the filter in shaping the acoustic output of their voice.

To address this question, we ran a production experiment to explore the resonance adjustment associated with pitch production with ultrasound tongue imaging. This technique allows us to directly visualize the actual articulatory strategies engaged by tongue movement in the singing task without needing to infer the filter function through acoustics, as resonance structures are difficult to track at higher F0s using linear predictive coding on audio data only. There are no relevant publicly available datasets to the authors' knowledge.

2. Methods

23 participants (15 female) aged between 18-35 participated in a production study using ultrasound tongue imaging technique. Participants were either non-singers or amateur singers who only sing for pleasure (unpaid). Most participants reported either no musical training, or singing experience in a choral or acapella setting, with five of the participants having also reported some solo experience. Participants were all fluent speakers of English, and we did not exclude participants based on their language backgrounds. All participants provided signed consent forms, and the experiment protocol was approved by the Institutional Review Board at the authors' university.

We chose five English monophthongs as the target vowels of this study, including [i, ɛ, æ, ɑ, u]. These five vowels were presented in English monosyllabic words in [b_t], [t_d/t], and [s_d] contexts, these contexts were chosen because (near-) minimal sets were possible to produce and because as they had a place articulation closer to the anterior part of the mouth, both for articulatory ease and to minimize coarticulatory effects between dorsal consonants and the target vowels. A sixth filler vowel [ɔ] was also included at the end of each set of vowels. A filler was required as speakers tend to lengthen the final syllable of each breath group, due to both prosodic and aerodynamic reasons. As such, the final vowel would not have been consistent in tempo and air pressure as the target vowels. The vowel [ɔ] was chosen as the filler because some speakers would not have contrasted [ɑ] and [ɔ]. Each participant was assigned five randomized sets of words (e.g. bot boot bat beat bet bought), which included all three consonant contexts. The order of the target vowels were randomized, and the filler vowel was always positioned at the end of each set of stimuli.

For each given set of vowels, participants were asked to sing all 6 words in order from the lowest pitch to the highest pitch of their comfortable pitch range, rising by semitone, resembling a vocal warm-up exercise. For example, a participant may sing "seed, sod, sued, said sad, sawed" at G3, then the same words at G#3, and again at A3 and so forth. To keep pitch accuracy, the experimenter played the corresponding note on an electric piano between each recording. The experiment took place in a sound-attenuated booth and was administered over Articulate Assistant Advanced (AAA). Ultrasound data were recorded using an EchoB 128 CEXT-1Z Compact External ultrasound system with a 128 element microconvex ultrasound transducer.

Ultrasound images were splined in AAA using the preexisting trained DeepLabCut models provided by Articulate Instruments Ltd. [22, 23] Audio recordings corresponding to the tongue imaging data were force aligned using the Penn Phonetics Lab Forced Aligner [24] to create corresponding time alignments for each recording. Based on these time alignments, we averaged the coordinates of the mid-50 percent of each vowels' entire duration, to get the most stable portions of the sung steady state vowels. With each token representing one sung vowel, a total of 11280 averaged tongue spline tokens entered the analysis across all participants.

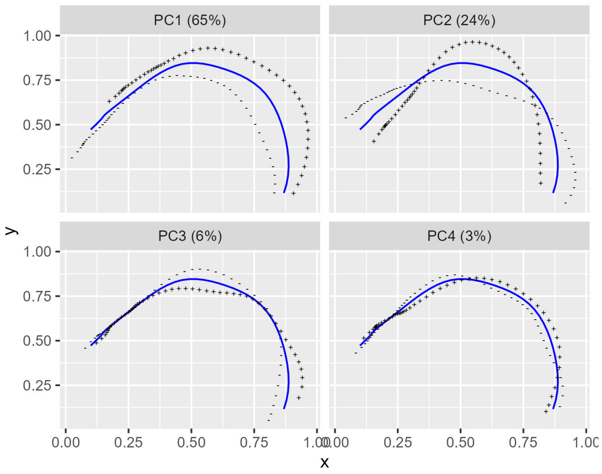

Figure 1: *Reconstructed tongue contours from the PCs of the x and y coordinates FPCA models plotted against each other (Tongue tip pointing left.)*

The entire shapes of the tongue contours were analyzed using functional principal component analysis (FPCA) with the *fda* package in R. In order to maximize interpretability as well as the amount of variation captured in the analysis, we ran two separate FPCAs, one modeling the x-coordinates only, and the other modeling the y-coordinates only. This allows us to regenerate the tongue contour when plotting results from one model against the other. The tongue splines were pre-processed using the smooth.basis() function to construct a functional data object by smoothing the tongue splines. To create the basis function required for defining the curves during data smoothing, we created a B-spline basis functional data object using the create.bspline.basis() function, since our tongue spline functions are non-periodic. We included 9 knots spaced equally apart, with the order of b-splines set at cubic splines. After preprocessing, we then applied FPCA to the two smoothed functional data objects, representing the x and y coordinates of our tongue contours.

3. Results

Figure 1 shows the first four FPCs and their corresponding factor scores. Positive factor values are indicated by the line consisting of (+) signs and negative factor values are indicated by the line consisting of (-) signs; the blue solid line is the average tongue position shown identically in all four facets. The amount of variation captured on the FPCA model for the x-axis were 76.0%, 15.6%, 4.4%, and 1.6% for principal components (PCs) one to four respectively, and the variation captured by the y-axis model were 53.5%, 32.4%, 6.8%, and 3.8% for PCs one to four respectively. The average variation captured is shown on the facet headers in figure 1. Since PC1 and PC2 only already account for 91.6% and 85.9% of the variability for the x and y coordinates respectively, we focus only on PC1 and PC2. Overall, figure 4 suggests that, tongue position variation of PC1 is related to backness and height, and PC2 is related to height.

Figure 2 is included to help better understand what the pos-

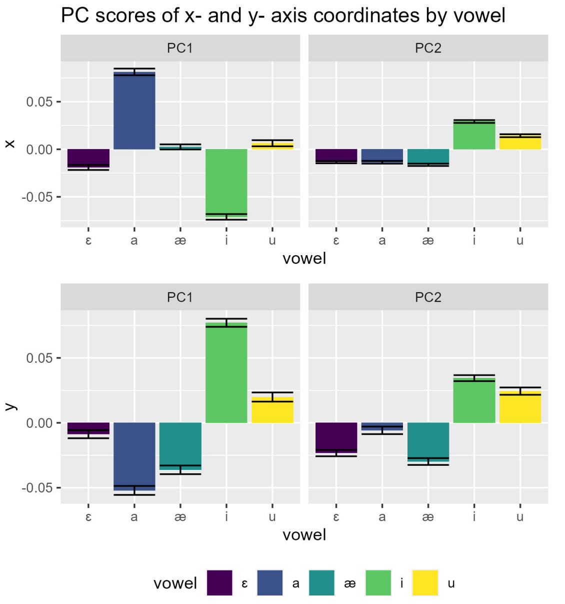

Figure 2: *PC scores of x-coordinates (top) and y-coordinates (bottom) by vowel.*

itive and negative factors entail. The figure indicates the mean FPC score of each vowel quality across all pitch ranges, showing how each vowel quality is related to the four FPCs. A positive score suggests that the vowel deviates from the mean tongue shape towards the positive line on a given axis. Conversely, a negative FPC score suggests that the vowel deviates from the mean tongue shape towards the negative line on a given axis.

Comparing figure 1 and figure 2 gives us an indication that PC1 of the x-axis coordinates is rough representative of vowel backness. The vowel [ɑ] appears to deviate from the mean with a positive score, such that it is closer to the line with a (+) sign, indicating backing. And both vowels [ɛ] and and [i] show negative PC scores, where [i] as the most front vowel shows more negative PC scores on the x-axis. Conversely, PC1 of the y-axis is a rough representation of vowel height. As can be seen, mid and low vowels [ɛ], [æ], and [ɑ] have a negative PC score, suggesting that they deviate from the mean towards the negative tongue contour on the y-axis, and the high vowels [u] and [i] have a positive PC score, suggesting that they deviate from the mean towards the positive tongue contour on the y-axis. Noteworthy is that PC2 for both x and y coordinates are also separated by vowel height, such that mid and low vowels [ɛ], [æ], and [ɑ] have a negative PC score, and high vowels [i] and [u] have a positive FPC score. Since the direction of the scores is congruent across x and y coordinates, this means that for FPC2, we may interpret that high vowels have an overall tongue shape closer to that of the positive line and mid or low vowels have an overall tongue shape closer to the negative line. This information is summarized in figure 3.

To address the question of whether speakers engage in vowel modification across their pitch range, we summarize PC scores by speakers' target pitch by semitone in figure 4. Target pitch 1 represents the lowest pitch that a speaker was comfortable singing that day, and target pitch 20 would represent 20 semitones above the lowest pitch of a given speaker. The minimum range a participant was willing to sing was 5 semitones, and the maximum range a participant was willing to sing was

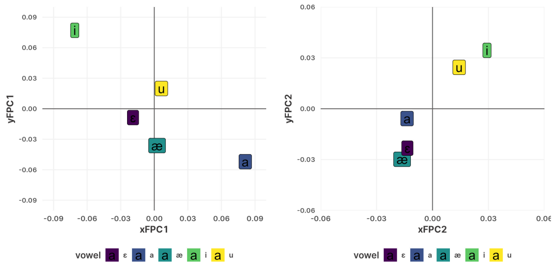

Figure 3: *PC1 and PC2 as rough representations of articulatory vowel spaces*

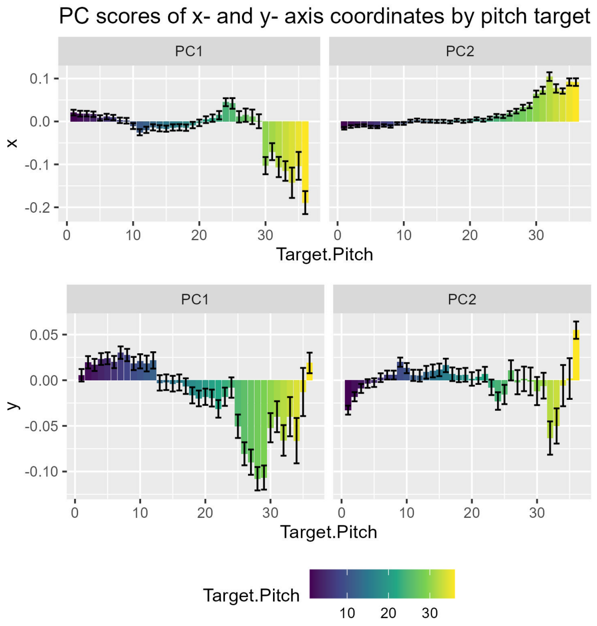

Figure 4: *PC scores of x- and y-coordinates by target pitch.*

36, with both the mean and the median range across all participants being 25 semitones.

Focusing on PC1, as the target pitch increases, both the x and y coordinates indicate a tendency towards a negative PC score. Referencing figure 1, a negative PC1 score for the x-axis means that the tongue contours deviate from the mean as the pitch increases to the maximum range closer to the negative line on the horizontal axis, such that the tongue root is more fronted than the grand mean. A negative PC1 score for the y-axis suggests that the tongue contour is also deviating from the mean towards the negative line on the verticle axis, suggesting that the tongue position is also lowering as pitch increases. In other words, participants tend to modify vowels to be both fronter and lower when they reach the higher end of the pitch range. As for PC2, while there does not appear to exist a clear flow in direction of the y-axis PC scores, the x-axis PCs tend to move towards a higher PC score as participants reach higher pitch targets, meaning that participants' tongue contour shape shifts towards the positive line, which is a high-vowel-like gesture. Since this concerns the horizontal axis, this does not mean participants' tongues are raising, instead it appears to be extra indication of tongue root fronting.

To statistically test whether participants' tongue position modification contribute to the production of pitch targets, we ran a linear mixed effects model using the lme4 package [25] to predict participants' pitch target. We included the main effects of FPC1 and FPC2 scores of the x coordinates, y coordinates, and target vowel quality, as well as two-way interaction effects between each of the FPC scores and vowel quality. The reference level for vowel was [ɑ]. We included participants as random intercepts but not any random slopes due to non-convergence.

Overall results show significant main effects of both the PC scores for x and y coordinates for both PC1 (X: β = -5.788, SE = 1.042, t = -5.555, $p < 0.0001$; Y: β = -5.396, SE = 1.070, t = -5.043, $p < 0.0001$) and PC2 (X: β = 12.990, SE = 2.368, t = 5.485, $p < 0.0001$; Y: β = 11.090, SE = 1.356, t = 8.180, $p < 0.0001$), suggesting that tongue shape on both dimensions do modulate by pitch target overall. The estimates for PC1 in both x and y coordinates are negative, and are positive for PC2, these observations are in line with our earlier observation from figure 4. We also observe significant interaction effects between all vowel qualities and PC1 of the x coordinate ([æ]: β = 6.279, SE = 1.624, t = 3.866, p = 0.0001, [ɛ]: β = 8.895, SE = 1.589, t = 5.598, $p < 0.0001$, [i]: β = 13.350, SE = 1.480, t = 9.025, $p < 0.0001$, [u]: β = 4.530, SE = 1.451, t = 3.122, p = 0.0018), and PC2 of the x coordinates ([æ]: β = 9.611, SE = 3.445, t = 2.790, p = 0.0053, [ɛ]: β = 11.560, SE = 3.455, t = 3.346, p = 0.0008, [i]: β = -7.305, SE = 3.186, t = -2.293, p = 0.0219, [u]: β = -14.740, SE = 3.158, t = -4.668, $p < 0.0001$). As for the y coordinates, the vowels [i] (β = -6.419, SE = 1.463, t = -4.386, $p < 0.0001$) and [u] (β = -2.967, SE = 1.330, t = -2.230, p = 0.0257) interacted significantly with PC1, and PC2 observes significant interaction effects with [æ] (β = 6.350, SE = 1.722, t = 3.688, p = 0.0002) and [ɛ] (β = 6.389, SE = 1.752, t = 3.646, p = 0.0003). Together, these results suggest that each individual vowel quality interacts with general vowel backness modification differently to different extents compared to the reference token.

To investigate whether there is a relationship between pitch range and tongue positional adjustments, and especially for participants with no singing experience, we ran a second linear mixed effects model isolating data only from participants who reported no history of singing solo, nor any history of singing in an ensemble setting like choirs or acapella groups. The subsetted participants' pitch ranges ranged from 5 semitones to 24 semitones. We predicted the pitch target with the main effects of the first two PC scores of the x and y coordinate models and two-way interaction effects between each of the PC score parameters with participants' maximum pitch range in a continuous scale. We exclude participants' pitch range as a main effect in the model as it would be redundant to predict pitch target by pitch range, since we are only interested in the interaction effects with the PCs. The model results show no significant main effects of any of the PC coordinates representing the x and y models, suggesting that the average participant without any vocal experience does not engage in vowel modification. When considering interacting effects between participants' maximum pitch range and the PC scores, the model indicates significant interaction effects between pitch range and PC1 scores for both x coordinate PC scores (β = -0.702, SE = 0.121, t = -5.811, $p < 0.0001$) and y coordinate PC scores (β = -0.280, SE = 0.121, t = -2.314, p = 0.021). The negative beta scores suggest that as participants with a larger pitch range sing higher pitches, their PC1 scores on both axes decrease, which is congruent with the direction of vowel modification we have observed – tongue lowering and tongue fronting. No significant interaction effects for neither x (β = 0.080, SE = 0.256, t = 0.313, p = 0.754) nor y (β = 0.196, SE = 0.170, t = 1.149, p = 0.251) coordinates' PC2 scores were observed. Together, the model results suggest that even among participants who reported neither training as a soloist nor experience in singing, participants who do know to modify their vocal tracts when completing the task tend also to be able to more effectively sing in wider pitch ranges.

4. Discussion

Professional singers engage in modification of their vocal tract space depending on the target pitch and especially when approaching higher pitches. While resonance tuning and vowel modification are prominent techniques used in the Western classical tradition, the support that the resonance space could give to the voice source has been acknowledged by singers of several traditions. We seek to ask in this study whether vocal tract adjustments by pitch is a technique learned only by singers to address acoustic goals, or whether it is part of speakers' knowledge of voice production for articulatory ease. We test this question by recruiting speakers who are not professional singers to engage in a singing task with ultrasound tongue imaging to visualize the mechanism behind vocal tract adjustments.

A main finding in our paper is that across the board, speakers who are not professional singers do engage in modification of vowels when targeting different pitch ranges. Namely, in this case, tongue fronting and lowering. We do not attempt to map this gesture uniquely to any specific known resonance tuning pattern in this paper since resonance tuning strategies could vary depending on many factors such as "singer proficiency, vowel, fundamental frequency (F0), vocal loudness, and phonation mode" [26] — not all of which are fully explored and discussed in this current paper given the space limitations. But we are able to conclude that the overall direction that speakers modify appears to be acoustically beneficial as both fronting and lowering gestures are ways to increase the resonance frequency of the lowest two resonances, where lowering increases the first resonance, and fronting increases the second resonance.

More importantly, our results from non-professional singers indicate that resonance modification is inherently required for more effective high-pitch production, and that is not a technique that is only acquired with extensive training. To validate this finding, we further excluded the participants who have some experience in singing (e.g., in choral or other ensemble settings), and tested the relationship between pitch target and tongue adjustments on complete layman speakers. Our results suggest that while the general trend among non-singers is that they do not make significant vocal tract adjustments when pursuing different parts of their pitch ranges, those who do have a larger pitch range tend to make vocal tract positional adjustments when approaching higher pitches. The choice of their tongue adjustment direction is also largely consistent with the rest of the relatively more experienced participants.

Together, our results suggests that vowel modification as a technique is not only a method for reaching acoustic goals driven by musical style, but instead is also important for articulatory ease for more effective voice production. These results are interesting as they provide direct evidence for how the filter structure may support voice source production, a relationship not often discussed when considering the linear source filter theory. To tease apart individual voice training vs general experiential effects in greater detail, alongside comparisons with professional singers and of different singing traditions will be a natural next step to this project.

5. References

[1] H. Hollien and J. F. Michel, "Vocal fry as a phonational register," *Journal of Speech and Hearing Research*, vol. 11, no. 3, pp. 600–604, 1968.

[2] R. J. Baken and R. F. Orlikoff, *Clinical Measurement of Speech and Voice*. San Diego: Singular Publishing, 2000.

[3] C. T. Herbst and B. H. Story, "Computer simulation of vocal tract resonance tuning strategies with respect to fundamental frequency and voice source spectral slope in singing," *The Journal of the Acoustical Society of America*, vol. 152, no. 6, pp. 3548–3561, 2022.

[4] G. Carlsson and J. Sundberg, "Formant frequency tuning in singing," *Journal of Voice*, vol. 6, no. 3, pp. 256–260, 1992.

[5] D. G. Miller and H. K. Schutte, "Formant tuning in a professional baritone," *Journal of Voice*, vol. 4, no. 3, pp. 231–237, 1990.

[6] J. Sundberg, "Formant technique in a professional female singer," *Acta Acustica united with Acustica*, vol. 32, no. 2, pp. 89–96, 1975.

[7] J. Sundberg, F. M. Lã, and B. P. Gill, "Formant tuning strategies in professional male opera singers," *Journal of Voice*, vol. 27, no. 3, pp. 278–288, 2013.

[8] M. Garnier, N. Henrich, L. Crevier-Buchman, C. Vincent, J. Smith, and J. Wolfe, "Glottal behavior in the high soprano range and the transition to the whistle register," *The Journal of the Acoustical Society of America*, vol. 131, no. 1, pp. 951–962, 2012.

[9] E. Joliveau, J. Smith, and J. Wolfe, "Tuning of vocal tract resonance by sopranos," *Nature*, vol. 427, no. 6970, pp. 116–116, 2004.

[10] M. P. Y. Chan and Y. Do, "Vowel modification (aggiustamento) in soprano voices," *Music & Science*, vol. 4, p. 20592043211055168, 2021.

[11] N. Henrich Bernardoni, J. Smith, and J. Wolfe, "Vocal tract resonances in singing: variation with laryngeal mechanism for male operatic singers in chest and falsetto registers," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 491–501, 2014.

[12] J. Sundberg, F. M. Lã, and B. P. Gill, "Professional male singers' formant tuning strategies for the vowel /a/," *Logopedics Phoniatrics Vocology*, vol. 36, no. 4, pp. 156–167, 2011.

[13] B. Coffin, *Coffin's overtones of bel canto: Phonetic basis of artistic singing: With 100 chromatic vowel-chart exercises*. Scarecrow Press, 1980.

[14] N. Henrich, M. Kiek, J. Smith, and J. Wolfe, "Resonance strategies used in bulgarian women's singing style: A pilot study," *Logopedics Phoniatrics Vocology*, vol. 32, no. 4, pp. 171–177, 2007.

[15] J. Schlömicher-Thier, D. Miller, H. Noé, and C. Herbst, "Yodeling—acoustic and physiologic properties," in *Poster at the 38th Annual Symposium of the Voice foundation*, 2009, pp. 3–7.

[16] J. Sundberg, B. Lindblom, and A.-M. Hefele, "Voice source, formant frequencies and vocal tract shape in overtone singing. a case study," *Logopedics Phoniatrics Vocology*, vol. 48, no. 2, pp. 75–87, 2023.

[17] M. E. Bestebreurtje and H. K. Schutte, "Resonance strategies for the belting style: results of a single female subject study," *Journal of voice*, vol. 14, no. 2, pp. 194–204, 2000.

[18] E. Björkner, "Musical theater and opera singing—why so different? a study of subglottal pressure, voice source, and formant frequency characteristics," *Journal of Voice*, vol. 22, no. 5, pp. 533–540, 2008.

[19] T. Bourne and M. Garnier, "Physiological and acoustic characteristics of the female music theater voice," *The Journal of the Acoustical Society of America*, vol. 131, no. 2, pp. 1586–1594, 2012.

[20] J. Sundberg, M. Thalén, and L. Popeil, "Substyles of belting: phonatory and resonatory characteristics," *Journal of Voice*, vol. 26, no. 1, pp. 44–50, 2012.

[21] D. Z. Borch and J. Sundberg, "Some phonatory and resonatory characteristics of the rock, pop, soul, and swedish dance band styles of singing," *Journal of Voice*, vol. 25, no. 5, pp. 532–537, 2011.

[22] A. Wrench and J. Balch-Tomes, "Beyond the edge: markerless pose estimation of speech articulators from ultrasound and camera images using deeplabcut," *Sensors*, vol. 22, no. 3, p. 1133, 2022.

[23] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge, "Deeplabcut: markerless pose estimation of user-defined body parts with deep learning," *Nature neuroscience*, vol. 21, no. 9, pp. 1281–1289, 2018.

[24] J. Yuan, M. Liberman *et al.*, "Speaker identification on the SCOTUS corpus," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3878, 2008.

[25] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[26] J. Sundberg, "Level and center frequency of the singer's formant," *Journal of voice*, vol. 15, no. 2, pp. 176–186, 2001.