2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct) | 979-8-3315-0691-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISMAR-Adjunct64951.2024.00034

Novel Sensing Methods for Vocal Technique Analysis: Evaluation on Electromyography and Ultrasonography

Kanyu Chen*
Keio University, Japan

Erwin Wu†
Tokyo Institute of Technology, Japan

Daichi Saito‡
Tokyo Institute of Technology, Japan

Yichen Peng§
Tokyo Institute of Technology, Japan

Chen-Chieh Liao ¶
Tokyo Institute of Technology, Japan

Akira Kato‖
Keio University, Japan

Hideki Koike **
Tokyo Institute of Technology, Japan

Kai Kunze ††
Keio University, Japan

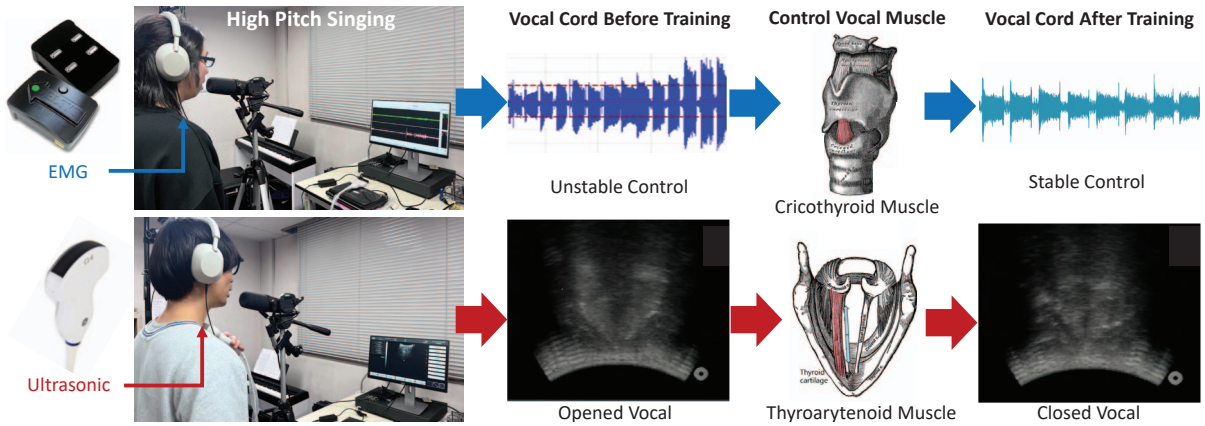

Figure 1: The system overview: (1) Showing a participant wearing EMG sensors to detect the activity of cricothyroid muscle. (2) Showing a participant using a B-type ultrasonography device to visualize the thyroarytenoid muscle.

Abstract

Controlling vocal cord muscles is crucial for vocal performance and training, yet it is challenging to measure. This paper introduces electromyography (EMG) and ultrasonography to detect vocal muscle activity and assess pitch training skills. A pre-experiment with 16 participants analyzed muscle control discrepancies among singers of different skill levels. A subsequent user study with 12 participants evaluated EMG and ultrasonography feedback effectiveness. Findings indicate that EMG offers better temporal stability representation, while ultrasonography provides intuitive visual feedback on vocal cord activity. Both methods show potential in enhancing vocal control, offering insights for designing effective and non-invasive vocal training systems.

Index Terms: human state, EMG measurement, ultrasonic sensing, acoustic interaction

1 Introduction

Recognizing and accurately controlling vocal cords is more important than merely improving performance in vocal training [13]. Current professional vocal training methods primarily rely on spectrograms for sound analysis [7] and laryngoscopy [12] for visualizing the state of the vocal cords. However, the spectrograms of recorded sound lack an intuitive understanding of vocal production. Although laryngoscopy is real-time instruction, it is too invasive and clinical. Thus, our study aims to address these limitations by providing a more intuitive and less invasive visualization method for vocal cords training.

Pitch control in voice production is determined by the tension of muscles [16], such as the thyroarytenoid and cricothyroid muscles (Fig 1). Common methods to measure muscle movements include electromyography (EMG) and ultrasonography [1, 5]. However, differences in vocal muscle control skills between novices and professionals are not well understood. Identifying these differences is also crucial for designing effective visualization methods for training.

Therefore, this paper addresses the following questions: (1) How can EMG and ultrasonography effectively evaluate vocal pitch performance? (2) How do users perceive and interpret the feedback from these sensing technologies?

This paper presents a pre-experiment using both EMG and ultrasonography to analyze vocal muscle activity related to pitch and control skill discrepancies among singers of different levels. We then evaluate and discuss the effectiveness and usability of both modalities in a vocal-training user study.

The contributions are summarized as follows: (1) a pre-experiment with 16 participants using both EMG and ultrasonography, providing an initial understanding of muscle control discrepancies among singers of varying skill levels; (2) an accessible approach to collect and evaluate vocal pitch using EMG and ultrasonography; (3) a user study with 12 novices assessing the effectiveness and usability of both feedback modalities in vocal training.

*e-mail: cady.cky@kmd.keio.ac.jp
†e-mail: wu.e.aa@vogue.cs.titech.ac.jp
‡e-mail: saito.d.ah@m.titech.ac.jp
§e-mail: peng.y.ag@m.titech.ac.jp
¶e-mail: liao.c.aa@m.titech.ac.jp
‖e-mail: kato@kmd.keio.ac.jp
**e-mail: koike@c.titech.ac.jp
††e-mail: kai@kmd.keio.ac.jp

2771-1110/24/$31.00 ©2024 IEEE
DOI 10.1109/ISMAR-Adjunct64951.2024.00034

2 Related Works

Sensing technology for vocal applications is enhancing the understanding and control of the human voice. Traditional methods [9] relied on subjective assessments, while recent EMG and ultrasonography advancements allow deeper exploration of vocal physiology in understanding the speech production.

EMG captures muscle activation patterns essential for vocal control [1], with high accuracy in recognizing Japanese vowels [8] and effectiveness in recognizing numerical speech commands [2]. Research has explored speech dynamics, including muscle engagement during vowel/sentence utterance [18], pitch and EMG for omohyoid detection [17], and face-neck muscle movements [10]. Wearable designs for the throat or chest, like custom collars [11], offer the potential for enhancing vocal performance.

Ultrasonography offers a non-invasive way to visualize vocal cord movements. It has shown diagnostic potential for vocal cord dysfunction [6] and has been used with deep learning for silent communication [5]. Yet exploration in vocal technique remains lacking.

3 Pre-Experiment: Vocal Cord Sensing

The primary goal of pre-experiment is to evaluate how to differentiate between beginners and experts in the activity of vocal muscles during singing using EMG and ultrasonography, and understand training feedback of design for users.

3.1 Experimental Setup

3.1.1 Data Collection

To avoid disruption from cables or attachments, we used two Delsys Trigno Wireless EMG sensors with 2-channel EMG, positioned on either side of the Adam's apple, capturing data at 2000Hz, downsampled to 30Hz (Fig 1). For ultrasonography, we used the CONTEC CMS600P2, with users holding the probe towards the vocal cords (Fig 1). The ultrasound captured video at 30 fps and 3.5 MHz. All data were aligned on the capture PC with a maximum 300ms delay, which did not significantly affect real-time feedback. The collected data include EMG signals, ultrasonography videos, and reference videos with audio during the collection. Detailed statistics of the pre-experiment data are provided in Table 1.

3.1.2 Participants

Based on the setup, we conducted a pre-experiment with 16 participants (6 female, 10 male, aged 21-33, mean=25.7) with varying levels of vocal training experience, recruited from two local institutes. Among the 16 participants, ten participants are beginners with little vocal training experience, and three participants are intermediate amateurs who have basic vocal knowledge. The remaining three participants are experts who experienced professional vocal training for more than 10 years.

3.2 Procedure

All participants are first given an introduction to the sensing devices and the task, followed by a 5-minute practice session to familiarize themselves with the sensing devices and the task. Regarding the task, to normalize the results and simplify the data processing, all participants are asked to do their best to sing a vocal scale, for example, from G3 to A4 according to the scientific pitch notation (SPN). They are guided to maintain a 2-second duration per pitch. An 80 bpm piano guidance is played in real-time for the participants as a reference, and the participants are asked to closely follow the piano notes. Each participant performs the task using the two sensing devices: one using the EMG and the other using ultrasonography, respectively. Each session is repeated four times, resulting in an eight-round data collection. The order of sessions is counterbalanced among each user to reduce any learning effects. After the collection, all participants are interviewed about vocal training and their impressions of the two sensing technologies.

As a result, the collected Vocal Cord Sensing (VCS) dataset consists of a total recording of more than 10K seconds. The details of the data can be found in Table 1.

4 Data Analysis

Since the raw data consists of noise and redundant information, we post-process the data and analyze the results from different perspectives, to obtain insights from the collected data.

4.1 Data processing

For the EMG data, since the focus is on the stability and controllability of the cricothyroid, we tried to extract the stability from the data. The raw EMG data are first denoised through a moving average filter (window size = 10ms), and then a Hilbert transform is performed to calculate envelopes of the signal [3]. This stability of muscle activity [4] s is then calculated as follows:

$$s = \frac{1}{N-1}\sum_{t=1}^{N-1} \|20log\frac{A_{t+1}}{A_t}\| \quad (1)$$

Here, A_t is the envelope value at timestep t. This equation is designed with reference to shimmer measurement in voice, which is frequently used in the field of acoustic analysis [15]. The calculated stability allows us to compare the differences between each subject.

The ultrasonography video must be quantified for analysis. Thus, we developed a landmark detector for tracking the vocal cords. We focus on five key points in the video: The start (connection) of two vocal cords, the end of the inner side of vocal cords, and the end of the outer side of vocal cords, as shown in Figure 2. These 5 points discern changes in the true vocal cord structure and cartilage position based on previous research [6] Since the shape of the vocal differs for every participant, it is difficult to train a robust detector directly. Thus, we manually annotate the keypoints on an initial frame for each session. These key points can be further tracked using a Lucas-Kanade-Tomasi (LKT) tracker. With the positional data from the five key points, we can easily compute the length of the true vocal cords (depicted in red in Figure 2) as follows:

$$L = \frac{1}{2} * \left(Dist\left(P_{vs}, \frac{P_{lv1}+P_{lv2}}{2}\right) + Dist\left(P_{vs}, \frac{P_{rv1}+P_{rv2}}{2}\right)\right) \quad (2)$$

4.2 Pre-Experiment Result

To understand the discrepancy in the muscle-control skills between amateurs and professionals, we analyze the muscle stability and vocal cord length included in the VCS dataset.

Firstly, the processed data is statistically analyzed using repeated measures of one-way ANOVA among the three groups of different level participants (beginners, intermediate amateurs, experts). A significant difference is revealed for both EMG ($F(1.449, 13.04) = 8.752$, $p = 0.0065$), and Ultrasonography ($(1.508, 12.07) = 183.3$, $p = 0.0001$). Post-hoc tests further confirmed significant differences between expert and beginner groups (EMG: $p = 0.0059$, US: $p = 0.001$), indicating a huge difference in level. Since most of the beginners cannot correctly sing the notes, we focus more on the data between the intermediate and expert-level participants in the following analysis.

As for the stability score of EMG, we pick up the common range (G3-G4) of the intermediate and expert groups to perform a deeper investigation. The results per pitch are shown in the upper figure of Figure 3. Despite an overall better temporal stability of the expert group, it is interesting to see an obvious difference in the higher pitches (F4 and G4), which indicates the ability to control vocal muscles in high-pitch sounds of the experts.

Table 1: Statistics of the VCS dataset.

Dataset	Sampling	Beginner (1-10)										Intermediate Amateurs (11-13)			Expert (14-16)			Total Size
Subject	-	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	16
Pitch Range (the number of pitch)	G2 - E6 (27)	F3 - F4 (14)	C3 - C4 (8)	F3 - G4 (9)	C3 - B4 (14)	G3 - D5 (12)	G3 - D5 (12)	D3 - C5 (14)	G3 - C5 (11)	F3 - D5 (13)	F3 - D5 (13)	F3 - E5 (14)	F2 - C5 (19)	G2 - E5 (20)	E3 = E6 (22)	D3 - E6 (22)	G2 - C6 (25)	G2 - E6
2-channel EMG Data	2000 hz/s	316s	253s	179s	383s	348s	277s	342s	130s	245s	329s	298s	249s	421s	139s	495s	524s	4928s
Ultrasonography Data	30 fps	273s	333s	264s	287s	217s	213s	280s	336s	232s	288s	333s	288s	368s	483s	583s	360s	5138s

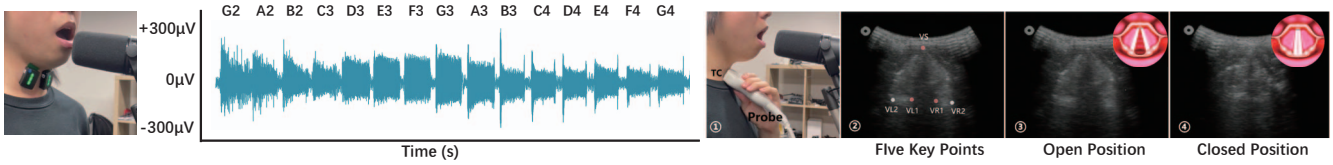

Figure 2: Left: A sample of raw EMG data. Right: A sample of raw Ultrasonography data accompanied by muscle/cartilage position annotations. The white and red points on the imaging of each pitch represent manually annotated keyframe data.

Regarding the vocal length data estimated from Ultrasonography, the same subjects are investigated as the EMG stability. The results are shown in the lower part of Figure 3. The results suggest a similar trend as the EMG data that the range of vocal cords is more stable for the experts. This shows how expert singers can precisely manage their vocal cords.

4.3 Insights

These results suggested that EMG stability and vocal cord length measured by ultrasonography are useful to indicate singer skill levels; thus, these sensors can potentially be used for the training. From the questionnaire, we found that the EMG sensor is easy to understand and provides a better representation of the temporal stability of the vocal cords. In addition, ultrasonography is a direct visual cue of the vocal cords, which can provide intuitive feedback on muscle activities.

5 User Study

To further evaluate the novice users' understanding of pitch performance and design future sensing feedback for vocal training, a user study for vocal pitch training is performed.

5.1 Participants

To investigate the training effects of visual cues from both above vocal cords sensing, 12 participants (6 male, 6 female, aged 24-34 years, mean=27.4) are recruited. Regarding singing (vocal training) experience, none of the participants have experienced professional training while all participants visit Karaoke (or similar activities) at least once a month.

5.2 Procedure

We prepare the three conditions: Baseline (training with only audio feedback), EMG (training with audio and EMG visual feedback), and Ultrasonography (training with Ultrasonography feedback).

Firstly, the participants warm up for minutes to be prepared for the vocal exercise. Next, we conducted the following procedure under all the conditions. The order of EMG and ultrasonography is counterbalanced. We measure the pre-training performance (Baseline/EMG/Ultrasonography) without a vocal guide. Users then practice for 5 minutes under one of the three conditions with a vocal guide featuring a brief demonstration of vocal muscle structure and expert's vocalis mechanism. After that, we measure the post-training performance without a vocal guide and ask questionnaires about the condition to the participants. After this repetition, an overall interview about the study regarding the participant's references, the evaluation, and suggestions for the system.

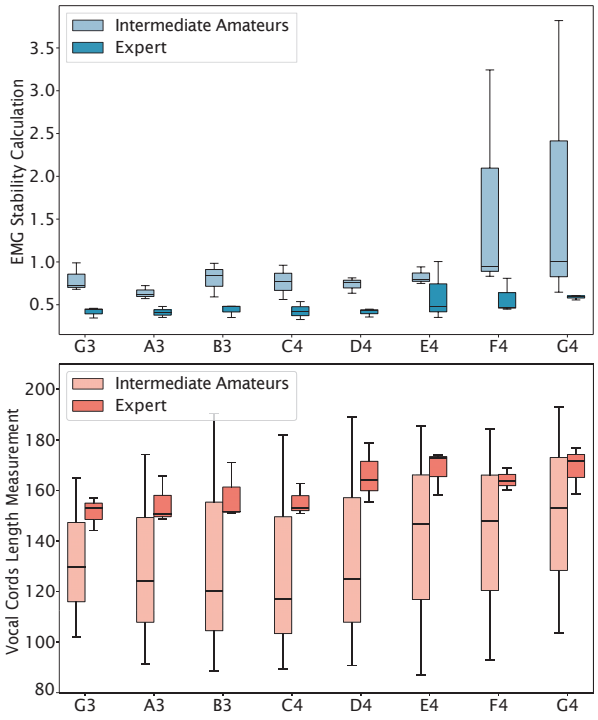

Figure 3: A set of a participant's raw EMG (upper), ultrasound (lower) imaging data (pitch ranged from G3-G4) spanning over one octave.

5.3 Results and Discussions

Based on our pre-experiment analysis methods, we conducted an EMG stability test and an ultrasonography vocal-length control test to evaluate participants' post-training vocal performance. Regarding self-reported perception, we used a subset of the Sense of Agency Scale (SoAS) [14] and interview.

5.3.1 EMG enhanced self-reported sense of control.

Analysis of the adapted SoAS questionnaire (Table 2) shows a heightened perception of controllability among participants using our proposed system compared to three alternative methodologies. Wilcoxon's signed-rank test revealed significant preferences for *EMG* over *Ultrasonography*, particularly in responses to *Q1*, *Q2*, and *Q5* ($W = 3, p = 0.0482$; $W = 3, p = 0.0451$; $W = 7, p =$

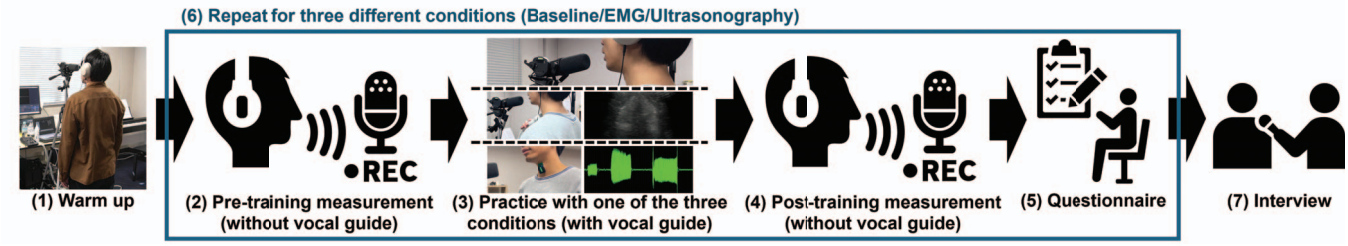

Figure 4: The procedure of a user study.

Table 2: The results of the pick-up SoAS questionnaire with the Wilxocon test. (*: $p < 0.05$, **: $p < 0.01$). Bolded numbers show the best scores.

Question	Baseline	Controllability	
		EMG	Ultrasonography
1. My actions just happen without my intention.↓	2.83 ±0.94	3.33 ±1.30	**2.41 ±1.16** *
2. My behavior is planned by me from the very beginning to the very end.↑	3.42 ±0.79	**3.83 ±0.83** *	2.92 ±1.24
3. I am completely responsible for everything that results from my actions.↑	3.92 ±0.67	**4 ±0.74**	3.42 ±1.0
4. My movements are automatic - my body simply makes them.↓	3.17 ±0.83	**2.25 ±0.97**	2.75 ±1.1
5. I am in full control of what I do.↑	3.08 ±0.90	**4.08 ±0.79** *	3 ±1.20

0.0125). The *EMG* approach demonstrated superior controllability compapred to the baseline and ultrasonography, likely due to the intuitive representation of EMG signals, and notably outperformed in *Q5* related to self-presentation confidence [14].

5.3.2 Over-attention may decreases EMG stability.

Increased self-reported perception may not align with muscle stability. Intriguingly, as depicted in Figure 5 (upper), the post-practice stability was generally lower than the pre-practice levels. This finding stands in contrast to the self-reported perceptions of increased controllability as indicated in the responses to the adapted SoAS questionnaire. Notably, this decrease in stability was statistically significant, especially for pitches *V3*, *C3*, and *E4*, with $W = 6, p = 0.048$, $W = 3, p = 0.049$, and $W = 14, p = 0.009$ respectively, as determined by the Wilcoxon's signed-rank test.

The decrease in EMG could be due to distraction as users mostly relied on keeping stable visual and cognitive demands associated with interpreting raw EMG. For instance, P8 noted, "The visual feedback draws most of my attention; I then try to control my vocal according to the visuals." Despite this, participants overall had a positive view of EMG training.

5.3.3 Ultrasonography increased mean vocal controlling

As illustrated in Figure 5 (lower), there was a noticeable increase in the mean vocal lengths for each pitch post-training, compared to the pre-training measurements. This improvement was particularly evident in pitches *A3*, *B3*, and *C4*, where statistical significance was established with $W = 3, p = 0.089$, $W = 7, p = 0.058$, and $W = 4, p = 0.049$ respectively, as ascertained through the Wilcoxon's signed-rank test.

This finding suggests improved vocal length control, aligning with VCS dataset trends where experts exhibited longer, more stable vocal cord lengths compared to intermediates. Our study shows experts had longer vocal lengths with less variability, indicating stable pitch control. Ultrasound-assisted training helped amateurs approach expert performance. "I noticed the teacher's vocal cords slowly widened from thin to wide when vocalizing, so I paid attention to reproducing this during practice", noted P7. However,

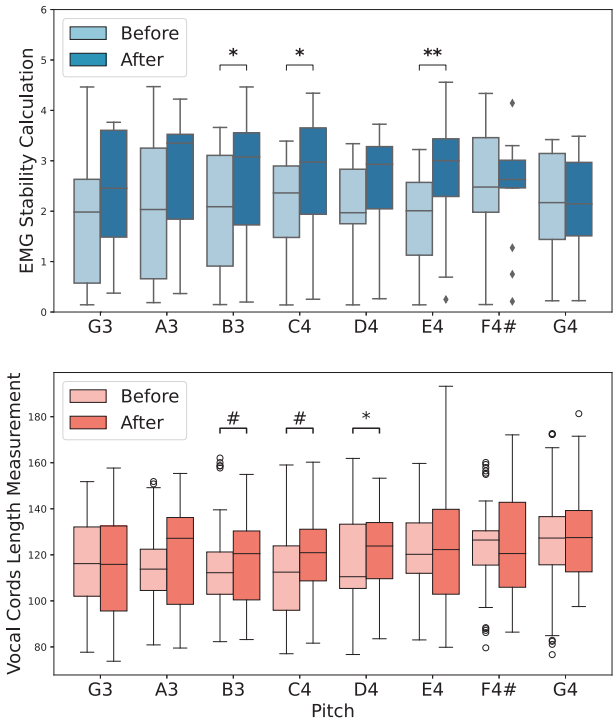

Figure 5: The results of the stability $s \downarrow$ test on the muscle activity (upper), and the results of the vocal length (lower). (#: $p < 0.1$, *: $p < 0.05$, **: $p < 0.01$)

some participants struggled with controlling vocal cord length, as P1 mentioned, "Good to see how vocalis moves, but hard to control." Usability issues with the ultrasound probe may have also affected results.

6 Conclusion and Future Works

In this paper, we proposed a novel approach using EMG and ultrasonography for vocal training. By collecting data from singers of varying skill levels and conducting a user study, we evaluated the effectiveness and usability of these modalities. EMG and ultrasonography effectively detected and differentiated vocal cord muscle activity, revealing significant differences between experts and intermediates. EMG highlighted muscle activation nuances, while ultrasonography provided insights into vocal cord length and dynamics. Despite perceived muscle control improvements with EMG, objective measurements showed lower post-practice stability, indicating a need for interface refinement. The result in the case using ultrasonography showed significant improvements in vocal length control, especially in specific pitch ranges, demonstrating the effectiveness of visual feedback. Future training systems should provide an intuitive interface, align with expert models, and incorporate extended training routines based on our findings.

Acknowledgments

This work is conducted under the IoT Accessibility Toolkit Project supported by JST Presto Grant Number JPMJPR2132, and is collaborate with Tokyo Institute of Technology under the fundings by JST Moonshot R&D Grant No. JPMJMS2012.

References

[1] F. Buchthal. Electromyography of intrinsic laryngeal muscles. *Quarterly Journal of Experimental Physiology and Cognate Medical Sciences: Translation and Integration*, 44(2):137–148, 1959. 1, 2

[2] A. D. Chan, K. Englehart, B. Hudgins, and D. F. Lovely. Myo-electric signals to augment speech recognition. *Medical and Biological Engineering and Computing*, 39:500–504, 2001. 2

[3] L. Cohen. *Time-frequency analysis*, vol. 778. Prentice hall New Jersey, 1995. 2

[4] M. Farrús, J. Hernando, and P. Ejarque. Jitter and shimmer measurements for speaker recognition. In *8th Annual Conference of the International Speech Communication Association; 2007 Aug. 27-31; Antwerp (Belgium).[place unknown]: ISCA; 2007. p. 778-81*. International Speech Communication Association (ISCA), 2007. 2

[5] N. Kimura, M. Kono, and J. Rekimoto. Sottovoce: An ultrasound imaging-based silent speech interaction using deep neural networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–11, 2019. 1, 2

[6] A. Kumar, C. Sinha, A. K. Singh, and U. K. Bhadani. Vocal cord dysfunction: ultrasonography-aided diagnosis during routine airway examination. *Saudi Journal of Anaesthesia*, 11(3):370–371, 2017. 2

[7] K. M. Luck, D. C. Lerman, W.-L. Wu, D. L. Dupuis, and L. A. Hussein. A comparison of written, vocal, and video feedback when training teachers. *Journal of Behavioral Education*, 27:124–144, 2018. 1

[8] H. Manabe, A. Hiraiwa, and T. Sugimura. Unvoiced speech recognition using emg-mime speech recognition. In *CHI'03 extended abstracts on Human factors in computing systems*, pp. 794–795, 2003. 2

[9] A. Naseth. Constructing the voice: Present and future considerations of vocal pedagogy. *The Choral Journal*, 53(2):39, 2012. 2

[10] C. Reed, A. McPherson, et al. Surface electromyography for direct vocal control. International Conference on New Interfaces for Musical Expression (NIME), 2020. 2

[11] C. N. Reed, S. Skach, P. Strohmeier, and A. P. McPherson. Singing knit: soft knit biosensing for augmenting vocal performances. In *Proceedings of the Augmented Humans International Conference 2022*, pp. 170–183, 2022. 2

[12] P. Song. Assessment of vocal cord function and voice disorders. *Principles and practice of interventional pulmonology*, pp. 137–149, 2013. 1

[13] J. Sundberg and T. D. Rossing. The Science of Singing Voice. *The Journal of the Acoustical Society of America*, 87(1):462–463, 01 1990. doi: 10.1121/1.399243 1

[14] A. Tapal, E. Oren, R. Dar, and B. Eitam. The sense of agency scale: A measure of consciously perceived control over one's mind, body, and the immediate environment. *Frontiers in Psychology*, 8:1552, 09 2017. doi: 10.3389/fpsyg.2017.01552 3, 4

[15] J. P. Teixeira and P. O. Fernandes. Acoustic analysis of vocal dysphonia. *Procedia Computer Science*, 64:466–473, 2015. 2

[16] I. R. Titze and B. H. Story. Rules for controlling low-dimensional vocal fold models with muscle activation. *The Journal of the Acoustical Society of America*, 112(3):1064–1076, 2002. 1

[17] J. M. Vojtech, M. D. Chan, B. Shiwani, S. H. Roy, J. T. Heaton, G. S. Meltzner, P. Contessa, G. De Luca, R. Patel, and J. C. Kline. Surface electromyography–based recognition, synthesis, and perception of prosodic subvocal speech. *Journal of Speech, Language, and Hearing Research*, 64(6S):2134–2153, 2021. 2

[18] M. Zhu, X. Wang, H. Deng, Y. He, H. Zhang, Z. Liu, S. Chen, M. Wang, and G. Li. Towards evaluating pitch-related phonation function in speech communication using high-density surface electromyography. *Frontiers in Neuroscience*, 16:941594, 2022. 2