Introduction of the VCSD Dataset: A Vocal Cords Dataset Using EMG and Ultrasonography for Singing Pitch Skill Recognition

Kanyu Chen
cady.cky@kmd.keio.ac.jp
Keio University
Institute of Science Tokyo
Tokyo, Japan

Erwin Wu
Chen-Chieh Liao
wu.e.aa@vogue.cs.titech.ac.jp
liao.c.aa@m.titech.ac.jp
Institute of Science Tokyo
Tokyo, Japan

Daichi Saito
Yichen Peng
saito.d.ah@m.titech.ac.jp
peng.y.ag@m.titech.ac.jp
Institute of Science Tokyo
Tokyo, Japan

Kato Akira
kato@kmd.keio.ac.jp
Keio University
Tokyo, Japan

Hideki Koike
koike@c.titech.ac.jp
Institute of Science Tokyo
Tokyo, Japan

Kai Kunze
kai@kmd.keio.ac.jp
Keio University
Tokyo, Japan

Abstract

We present the Vocal Cords Sensing Dataset (VCSD), a multimodal dataset for analyzing singing pitch skills using surface electromyography (EMG) and ultrasonography (UI). The dataset includes over three hours of recordings from 16 participants with varied singing experience, capturing EMG signals and ultrasound vocal fold motion. VCSD supports the recognition of vocal pitch control and muscle coordination during the pitch singing. Initial analysis shows significant differences in EMG stability and vocal fold dynamics between novice and expert singers. This dataset enables future research in vocal training, physiological sensing, and interactive feedback design.

CCS Concepts

• **Human-centered computing** → *Ubiquitous and mobile computing design and evaluation methods*; • **General and reference** → **Measurement**; • **Applied computing** → *Bioinformatics*.

Keywords

VCSD dataset, electromyography, ultrasonography, singing pitch recognition

ACM Reference Format:
Kanyu Chen, Erwin Wu, Chen-Chieh Liao, Daichi Saito, Yichen Peng, Kato Akira, Hideki Koike, and Kai Kunze. 2025. Introduction of the VCSD Dataset: A Vocal Cords Dataset Using EMG and Ultrasonography for Singing Pitch Skill Recognition. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion '25), October 12–16, 2025, Espoo, Finland*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3714394.3754407

UbiComp Companion '25, October 12–16, 2025, Espoo, Finland

ACM ISBN 979-8-4007-1477-1/2025/10
https://doi.org/10.1145/3714394.3754407

1 Introduction

Vocal training, particularly for techniques such as belting [5], requires fine-grained control over pitch and vocal fold coordination. These tasks engage intrinsic laryngeal muscles such as the thyroarytenoid and cricothyroid under highly dynamic conditions [17, 19]. However, students often struggle to understand the muscular basis of vocal production, and traditional feedback—typically through auditory instruction, spectrogram review, or invasive laryngoscopy [15]—is often limited to instructor-guided or static environments. Subjective assessments by instructors can vary in precision [9], and opportunities for self-reflection on muscle use are scarce, particularly during rehearsals or live stage practice.

Wearable sensing technology now enables non-technical users to access physiological feedback in natural environments [2, 10, 12, 13, 20]. Prior research shows that electromyography (EMG) and Singing Power Ratio (SPR) metrics can differentiate between novice and expert singers [14, 18]. Electromyography (EMG) and ultrasonography (UI) are methods to measure muscle movements [1, 7, 8]. The tension of muscles [17], such as the thyroarytenoid and cricothyroid muscles, allows pitch control in voice production. Building on this, biofeedback tools have potential to support reflective learning and independent skill acquisition in vocal training. However, existing datasets for this purpose remain scarce.

To address this gap, we present the Vocal Cords Sensing Dataset (VCSD), a multimodal dataset designed for singing pitch skill recognition and vocal muscle sensing. VCSD combines surface electromyography (EMG) and ultrasonography (UI) to capture the vocal muscle activity and vocal fold dynamics of singers with different skill levels. This dataset enables enables the development of new methods for vocal skill recognition, feedback, and bio-sensing-based interaction [3].

2 Sensors and Data Collection

The primary goal of this study was an exploratory analysis of the activity of vocal muscles during singing using EMG and UI to investigate whether these sensing technologies can capture different levels of proficiency in singing [3]. For this, we collected a Vocal Cord Sensing Dataset (VCSD).

Dataset	Sampling	Novices (1-10)										Intermediate Amatuers (11-13)			Experts (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

Table 1: Statistics of the VCSD datasets: EMG, UI, and pitch range for 16 users across skill levels (Novice, Intermediate, Expert)

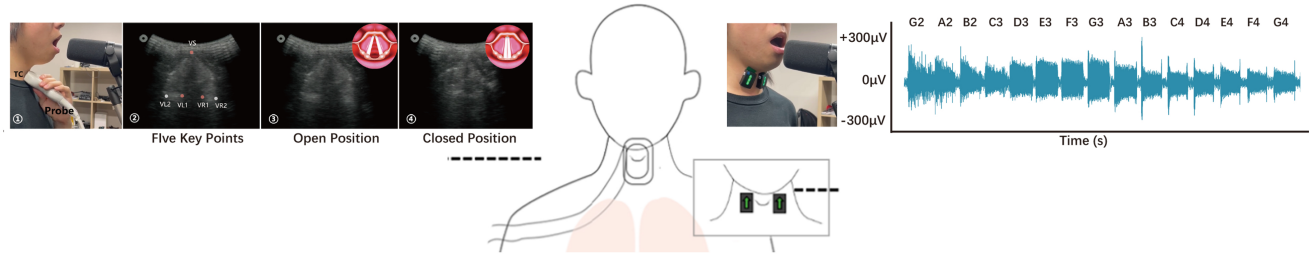

Figure 1: Left: EMG sensor placement and visualization of a sample of raw EMG data accompanied by muscle/cartilage position annotations. Right: Positioning of the ultrasonography probe and sample of raw ultrasound imaging data accompanied by muscle/cartilage position annotations.

2.1 Devices and Setup

We collected two types of data to directly measure muscle activity across various pitches: EMG and UI. A camera microphone collected audio reference.

EMG Setup: To capture EMG, we used Delsys Trigno Wireless EMG sensors to measure vocal muscle activity. The sensor was positioned between the Adam's apple and the first bone below. The data was captured at 2000Hz and down-sampled to 30Hz for analysis.

Ultrasonography: The CONTEC CMS600P2 Digital B-Ultrasound Diagnostic System was employed. Participants held the probe at the front of the larynx, capturing images at 3.5 MHz and video at 30 fps. Ultrasonography was used only in the novice study for visual feedback on vocal muscle activity. Due to its bulkiness, this method was not used in the professionals study.

2.2 Dataset Description

The VCSD dataset comprises over three hours of synchronized multimodal recordings collected during structured vocal exercises. It includes two-channel electromyography (EMG) signals sampled at 2000 Hz, ultrasonography (UI) videos captured at 30 frames per second, and synchronized audio-video recordings of participants for reference. These recordings were obtained under controlled pitch-singing tasks designed to elicit measurable differences in vocal performance. An overview of the data structure is provided in Table 1, and representative sensor placements along with raw signal examples are shown in Figure 1.

2.3 Participants & Recruitment

We recruited 16 participants (6 female, 10 male, aged 21-33, mean=25.7) from two local institutes. Ten participants were beginners with little vocal training experience, and three participants were intermediate amateurs who had basic vocal knowledge. The remaining three participants were experts who experienced professional vocal training for more than 10 years. Novice and experienced participants were provided a 1000 yen gift card per hour as reimbursement for their participation in the study. Expert participants were reimbursed with 10,000 yen as compensation for their time and effort. The collection process was approved by the IRB of the local institutes. All data collected within the scope of this study are anonymous and are only released with the approval of the participants.

2.4 Study Procedure

1) Welcome & Familiarization: Before data collection, the participants were informed about their rights and the implications of their participation. When agreeing to participation and use of their data, they signed a consent form. Then, an initial questionnaire was given to collect the participants' demographics and their vocal training experience. Next, participants were introduced to the sensing devices and the task, followed by a 5-minute practice session to familiarize themselves with the sensing devices and the task.

2) Singing Task: The participants were asked to do their best to sing a vocal scale, for example, ranging from G3 to A4 according to the scientific pitch notation (SPN). To support them, we played an 80 bpm piano guidance sound of the scale in real-time as a reference. The participants were asked to closely follow this reference and to maintain a 2-second duration per pitch. Each participant performed

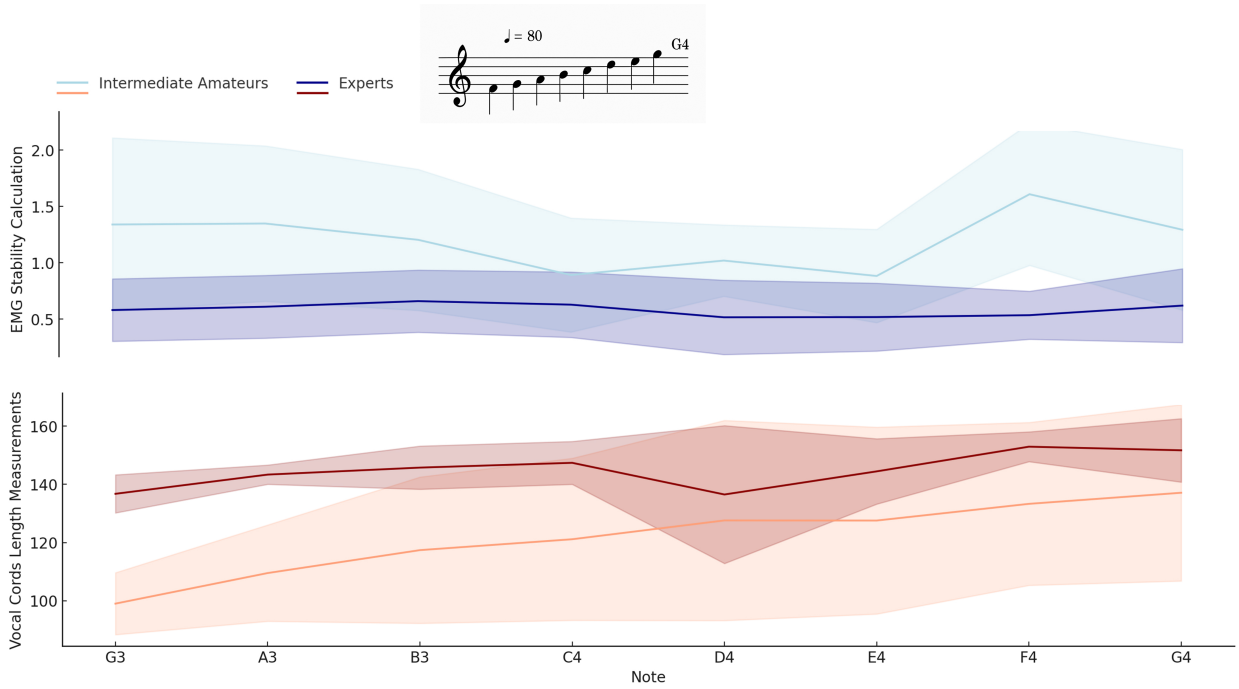

Figure 2: Comparison of the stability calculation derived from EMG data and the measured vocal cord length via ultrasonography across a pitch range from G3 to G4. Upper: a set of a participant's raw EMG imaging data; Bottom: a set of a participant's ultrasound imaging data, each spanning over one octave.

the task using the two sensing devices EMG and UI, respectively. Each session was repeated four times, resulting in an eight-round data collection. The order of sessions was counterbalanced among each user to reduce any learning effects.

3) Exit Interview: Finally, the participants were interviewed about vocal training and their impressions of the two sensing technologies.

3 Feature Extraction

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

3.1 Data Processing

EMG. Since our focus was on the stability and controllability of the cricothyroid muscle, we tried to extract stability information from the data. The raw data was first denoised through a moving average filter (window size = 10ms), then a Hilbert transform was performed to calculate envelopes of the signal [4, 11]. The stability of muscle activity [6] s was then calculated as follows:

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

A_t denotes the previous envelope value of the filtered EMG at timestep t. This equation is designed with reference to shimmer measurement in voice, which is frequently used in the field of acoustic analysis [16]. Dividing among envelopes allows calculating the stability of each pitch and comparing the differences in scales between each participant.

UI. The UI video had to be quantified for analysis. Therefore, we developed a landmark detector for tracking the vocal cord muscle from the ultrasonic images. Here, we focused on five important key points in the video: start points (connection) of two vocal cords, ends of the inner side of vocal cords, and the end of the outer side of vocal cords, as shown in Figure 1. These five points discern changes in the true vocal cord structure and cartilage position based on previous research [8]. Since the shape of the vocal cord differs for each participant, we manually annotated the key points on an initial frame for each session. These key points were further tracked using a Kanade-Lucas-Tomasi (KLT) tracker. With the positional data from the five key points, we computed the length of the true vocal cords (depicted in red in Figure 1) 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 Benchmark Analysis

First, to assess the ability of EMG and UI signals to distinguish between levels of vocal expertise, we conducted a repeated measures ANOVA among the three groups of different level participants (beginners, intermediate amateurs, experts). The ANOVA revealed significant differences for EMG ($F(1.449, 13.04) = 8.752$, $p =$

0.0065), and UI ($F(1.508, 12.07) = 183.3$, $p = 0.0001$). Post-hoc tests further confirmed significant differences between expert and beginner groups (EMG: $p = 0.0059$, UI: $p = 0.001$). Since most beginners cannot "correctly" sing the notes, we focused on the data between the intermediate and expert-level participants in the following analysis.

EMG. For the stability score, we picked 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 2. Despite an overall better temporal stability of the expert group, we found 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.

UI. For the vocal length data estimated from the UI, the same participants were investigated as for the EMG stability. The results suggest a similar trend as the EMG data that the range of vocal cords is more stable for the experts (see bottom part of Figure 2). Even though the UI does not provide any temporal information, it shows that expert singers can more precisely manage their vocal cords.

4.1 Skill Differentiation Insights

The analysis revealed clear physiological distinctions between skill levels. Experts exhibited lower EMG stability scores across all pitches, with particularly notable differences in higher notes such as F4 and G4. This suggests that expert singers possess more refined neuromuscular control during high-pitch vocalizations. Similarly, ultrasonography data showed that experts maintained longer and more consistent vocal fold lengths throughout the task, indicating a greater degree of anatomical control over vocal fold mechanics. These findings demonstrate that both EMG and UI are capable of capturing meaningful physiological markers associated with singing proficiency.

5 Dataset Availability

We plan to publicly release the VCSD dataset in CSV and video formats to support further research in vocal sensing and skill analysis. The dataset package will include time-series data from the EMG and annotated ultrasound landmarks represented as (x, y) coordinates, pitch labels, timestamps, and detailed session metadata. All data have been anonymized and are shared with participant consent under institutional ethics approval. The dataset will be hosted on a public osf research repository: https://osf.io/rbqcv/?view_only=3821784e313a4533a39c956a3c1dae6d.

6 Discussion

The VCSD dataset provides rich multimodal data for analyzing vocal performance. While our benchmarks demonstrate its capability in distinguishing singing skill levels, further work is needed to develop robust real-time feedback algorithms and personalized training interfaces.

Limitations include the relatively small number of expert participants and the task-constrained pitch range. Future extensions could incorporate diverse vocal styles, longer phrases, and additional modalities such as airflow or video-based facial tracking.

7 Conclusion

We presented the VCSD, a novel multimodal dataset that combines surface EMG and ultrasonography to capture vocal muscle activity during singing. By including participants across multiple skill levels, VCSD enables the development of methods for vocal skill assessment, training feedback, and physiological signal-based interaction. Initial benchmarks confirm the dataset's ability to differentiate singing proficiency through sensor-derived metrics, highlighting its value for the Ubicomp and HCI communities.

Acknowledgments

We thank all participants and institutional collaborators for their support. This work was funded in part by JST SPRING H09GQ24152, and 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] Fritz Buchthal. 1959. Electromyography of intrinsic laryngeal muscles. *Quarterly Journal of Experimental Physiology and Cognate Medical Sciences: Translation and Integration* 44, 2 (1959), 137–148.

[2] Kanyu Chen, Emiko Kamiyama, Ruiteng Li, Yichen Peng, Daichi Saito, Erwin Wu, Hideki Koike, and Akira Kato. 2024. Phantom Audition: Using the Visualization of Electromyography and Vocal Metrics as Tools in Singing Training. In *SIGGRAPH Asia 2024 Posters (SA '24)*. Association for Computing Machinery, New York, NY, USA, Article 113, 2 pages. https://doi.org/10.1145/3681756.3697908

[3] Kanyu Chen, Erwin Wu, Daichi Saito, Yichen Peng, Chen-Chieh Liao, Akira Kato, Hideki Koike, and Kai Kunze. 2024. Novel Sensing Methods for Vocal Technique Analysis: Evaluation on Electromyography and Ultrasonography. In *2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 121–125.

[4] Leon Cohen. 1995. *Time-frequency analysis*. Vol. 778. Prentice hall New Jersey.

[5] Jo Estill. 1988. Belting and classic voice quality: some physiological differences. *Medical problems of performing artists* 3, 1 (1988), 37–43.

[6] Mireia Farrús, Javier Hernando, and Pascual Ejarque. 2007. 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).

[7] Naoki Kimura, Michinari Kono, and Jun Rekimoto. 2019. 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*. 1–11.

[8] Amarjeet Kumar, Chandni Sinha, Akhilesh Kumar Singh, and Umesh Kumar Bhadani. 2017. Vocal cord dysfunction: Ultrasonography-aided diagnosis during routine airway examination. *Saudi Journal of Anaesthesia* 11, 3 (2017), 370–371.

[9] Pauline Larrouy-Maestri, Yohana Lévêque, Daniele Schön, Antoine Giovanni, and Dominique Morsomme. 2013. The evaluation of singing voice accuracy: A comparison between subjective and objective methods. *Journal of Voice* 27, 2 (2013), 259–e1.

[10] Takuro Nakao, Yun Suen Pai, Megumi Isogai, Hideaki Kimata, and Kai Kunze. 2018. Make-a-face: a hands-free, non-intrusive device for tongue/mouth/cheek input using EMG. In *ACM SIGGRAPH 2018 Posters*. 1–2.

[11] Alan V Oppenheim. 1999. *Discrete-time signal processing*. Pearson Education India.

[12] Yun Suen Pai, Tilman Dingler, and Kai Kunze. 2019. Assessing hands-free interactions for VR using eye gaze and electromyography. *Virtual Reality* 23 (2019), 119–131.

[13] Rakesh Patibanda, Nathan Arthur Semertzidis, Michaela Vranic-Peters, Joseph Nathan La Delfa, Josh Andres, Mehmet Aydin Baytaş, Anna Lisa Martin-Niedecken, Paul Strohmeier, Bruno Fruchard, Sang-won Leigh, et al. 2020. Motor memory in HCI. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[14] Viggo Pettersen, Kåre Bjørkøy, Hans Torp, and Rolf Harald Westgaard. [n. d.]. Neck and Shoulder Muscle Activity and Thorax Movement in Singing and Speaking Tasks with Variation in Vocal Loudness and Pitch. 19, 4 ([n. d.]), 623–634. https://doi.org/10.1016/j.jvoice.2004.08.007

[15] Phillip Song. 2013. Assessment of vocal cord function and voice disorders. *Principles and practice of interventional pulmonology* (2013), 137–149.

[16] João Paulo Teixeira and Paula Odete Fernandes. 2015. Acoustic analysis of vocal dysphonia. *Procedia Computer Science* 64 (2015), 466–473.

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

[18] M Usha, YV Geetha, and YS Darshan. 2017. Objective identification of prepubertal female singers and non-singers by singing power ratio using matlab. *Journal of Voice* 31, 2 (2017), 157–160.

[19] William Vennard. 1968. *Singing: the mechanism and the technic*. Carl Fischer, LLC.

[20] Jennifer M Vojtech, Michael D Chan, Bhawna Shiwani, Serge H Roy, James T Heaton, Geoffrey S Meltzner, Paola Contessa, Gianluca De Luca, Rupal Patel, and Joshua C Kline. 2021. Surface electromyography–based recognition, synthesis, and perception of prosodic subvocal speech. *Journal of Speech, Language, and Hearing Research* 64, 6S (2021), 2134–2153.