Effects on Voice Quality of Thyroidectomy: A Qualitative and Quantitative Study Using Voice Maps

Huanchen Cai, Sten Ternström, Philippe Chaffanjon, Nathalie Henrich Bernardoni

Effects on Voice Quality of Thyroidectomy: A Qualitative and Quantitative Study Using Voice Maps

[*]Huanchen Cai, [*]Sten Ternström, [†,‡]Philippe Chaffanjon, and [†]Nathalie Henrich Bernardoni, *Stockholm, Sweden, and †‡Grenoble, France*

Summary: Objectives. This study aims to explore the effects of thyroidectomy—a surgical intervention involving the removal of the thyroid gland—on voice quality, as represented by acoustic and electroglottographic measures. Given the thyroid gland's proximity to the inferior and superior laryngeal nerves, thyroidectomy carries a potential risk of affecting vocal function. While earlier studies have documented effects on the voice range, few studies have looked at voice quality after thyroidectomy. Since voice quality effects could manifest in many ways, that a priori are unknown, we wish to apply an exploratory approach that collects many data points from several metrics.

Methods. A voice-mapping analysis paradigm was applied retrospectively on a corpus of spoken and sung sentences produced by patients who had thyroid surgery. Voice quality changes were assessed objectively for 57 patients prior to surgery and 2 months after surgery, by making comparative voice maps, pre- and post-intervention, of six acoustic and electroglottographic (EGG) metrics.

Results. After thyroidectomy, statistically significant changes consistent with a worsening of voice quality were observed in most metrics. For all individual metrics, however, the effect sizes were too small to be clinically relevant. Statistical clustering of the metrics helped to clarify the nature of these changes. While partial thyroidectomy demonstrated greater uniformity than did total thyroidectomy, the type of perioperative damage had no discernible impact on voice quality.

Conclusions. Changes in voice quality after thyroidectomy were related mostly to increased phonatory instability in both the acoustic and EGG metrics. Clustered voice metrics exhibited a higher correlation to voice complaints than did individual voice metrics.

Key Words: Thyroidectomy–Voice quality–EGG–Classification–Voice mapping.

INTRODUCTION

Thyroidectomy, the surgical removal of the thyroid gland, is a common procedure for patients with thyroid cancer or other thyroid-related conditions, with over 100,000 thyroid operations performed annually in US,[1] about 45,000 in France, 60,000 in Germany, and 4000 in Switzerland.[2] A thyroidectomy may be the intervention chosen for diagnoses such as thyroid cancer, goiter, hyperthyroidism, thyroid nodules, and others. While the procedure is generally safe, there is a risk of damage, for example, to the nerves mediating the muscular control of the larynx.[3] It is well known that the ensuing effect on voice quality and voice control may be severe. The change in voice range is the major complaint that patients report after the operation.[4–7] However, even when the voice range is not severely impacted, patients sometimes complain of experiencing other qualitative changes to their voice. A few patients even report experiencing a deteriorated voice when they essentially sound better acoustically. Not many studies of such quality changes have been reported.[8] In the present retrospective study, we compare some voice signal attributes of thyroidectomy patients pre- and post-intervention, in terms of voice quality rather than voice range, using the voice-mapping paradigm and statistical clustering of several acoustic and electroglottographic (EGG) metrics.[9] By systematically mapping these metrics across a relevant vocal range, we look for signs of any previously undescribed alterations in voice quality characteristics. The quantitative data derived from this analysis are then analyzed for possible correlations with self-reported voice handicap indexes (VHIs) and specific types of thyroidectomies, in the hope of improving our understanding of postintervention voice quality changes.

Overview of thyroidectomy

Thyroidectomy is a surgical procedure that can lead to various changes in vocal characteristics. Voice function changes after thyroidectomy could be an effect of laryngeal edema, vocal fold bowing, orotracheal intubation trauma, extra laryngeal strap muscles damage or temporary malfunction of these muscles, and laryngotracheal fixation.[4,10–13] While certain surgical advancements, such as the minimally invasive parathyroid surgery, claim to preserve voice quality,[14] others, like total thyroidectomy (TT), have shown consistent evidence of voice quality degradation, even in the absence of vocal fold

Accepted for publication March 12, 2024.

A summary of this investigation was presented at the Annual Symposium on the Care of the Professional Voice, June 3, 2023.

Huanchen Cai's work is funded by the China Scholarship Council, grant number 202006010113.

From the *Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden; †University of Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, Grenoble, France; and the ‡Medical School, Université Grenoble Alpes, Grenoble, France.

Address correspondence and reprint requests to Huanchen Cai, Division of Speech, Music and Hearing, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden. E-mail: huanchen@kth.se

Journal of Voice, Vol xx, No xx, pp. xxx–xxx

0892-1997

https://doi.org/10.1016/j.jvoice.2024.03.012

paralysis.[15] Veldova[16] highlighted that voice disorders can occur even without recurrent laryngeal nerve (RLN) paresis after TT, suggesting that additional causes other than recurrent nerve paralysis may contribute to voice damage. Several studies have also shown that even when the laryngeal nerves are well preserved, complaints of voice changes and damage still occur.[17,18]

The diagnosis of injuries is monitored by intraoperative nerve monitoring (IONM). IONM has been advocated with the goal of reducing the rate of RLN injury. Although its routine use remains controversial,[1] it could potentially assist in the identification, dissection, and prediction of postoperative function of the RLN.[19]

Continuous stimulation could alert the surgeon to an impending nerve injury earlier than intermittent stimulation. However, continuous IONM has the disadvantages of potentially causing vagal neurapraxia (if the electrode dislodges) or hemodynamic instability (eg, cardiac arrest) secondary to increased parasympathetic (vagal) tone.[20] It also requires dissection of structures (eg, vagus nerve) outside of the typical operative field, although a transcutaneous method of vagal stimulation for IONM during endocrine surgery has been reported.[21]

Most studies examining the utility of IONM have been observational; the few randomized trials were small and underpowered. A 2019 Cochrane systematic review and meta-analysis of five trials found no advantage or disadvantage for IONM in either permanent (relative risk 0.77, 95% CI 0.33–1.77) or transient RLN palsy (relative risk 1.25; 95% CI 0.45–3.47), when compared with visual nerve identification.[22] A 2022 meta-analysis of eight trials similarly found no improvement with IONM in either permanent (IONM 0.5% vs visual nerve identification 0.6%, $P = 0.57$) or transient (1.5 vs 2.1%, $P = 0.11$) RLN injury, although there was a nonstatistically significant reduction in overall RLN injury (3.2 vs 2.3%, $P = 0.069$).[23]

Existing data do not support that the routine use of IONM reduces the incidence of RLN injury.[24–31] IONM, however, may be beneficial in procedures that are either high-risk or performed by low-volume surgeons.[24,32–34] High-risk procedures in this context typically include reoperations and those performed for thyroid cancers or goiters (retrosternal or toxic).

Impact on voice function changes

Numerous studies have shown that patients may experience a decrease in fundamental frequency, a higher jitter and shimmer, and a decrease in maximum phonation time after undergoing thyroidectomy.[10,35,36]

Similarly, a higher STD (standard deviation of fundamental frequency), a higher VTI (Voice Turbulence Index), and higher VHI (Voice Handicap Index) after surgeries have been observed.[37] Debruyne et al[38] examined the spoken voice quality of 47 patients preoperatively (1 week) and postoperatively on day 4. Acoustic parameters studied included fundamental frequency (f_o), frequency and intensity disturbances, harmonic prominence, H1-H2 difference, and spectral slope. Similarly, Van Lierde et al investigated vocal range and acoustic measurements using vowel /a/ sounds, with findings showing consistent decrease of f_o of the highest frequency, the highest intensity, and the Dysphonia Severity Index postoperatively at 3 months.[39]

However, the severity of these effects can vary greatly depending on the individual patient and the specifics of the surgery.[4,40–42] The impairment of cervical muscles or nerves resulting from intubation or surgical procedures can detrimentally affect voice quality, often manifested by a reduction in fundamental frequency. In cases of thyroidectomies where laryngeal nerve injury is absent, there may still be notable impacts on vocal capabilities.[42] Research indicates that total thyroidectomies tend to yield more pronounced disturbances in voice quality compared to partial thyroidectomies.[43] Heather et al discussed the effects of laryngeal cancer on voice quality.[44] Though vocal nodules, a common diagnosis, do not exhibit statistical difference in varying vocal nodule sizes,[45] hypothyroidism as a thyroid hormone deficiency has been identified as affecting human speech and voice.[46] Symptoms commonly observed include fluctuations in fundamental frequency and increased prevalence of voice disorders. Overall, the diagnosis of the underlying condition, the type of surgical intervention, and the extent of muscular or neural damage all contribute to variations in voice capability and quality.

Voice range considerations

The measurements of voice quality in the cited studies were typically based on extraction of only one or a few segments of voice. As reported by several studies,[47,48] the f_o and SPL (sound pressure level) can have a substantial influence on other metrics and indicators, though there is interindividual variability. Unless the elicited f_o and SPL are accounted for, such changes may mask the actual effects of the surgical and/or therapeutic interventions.[49] By making comparisons only at f_o and SPL that are matched pre- and post-intervention, such bias can be canceled. The matching is achieved in the areas that overlap each other, pre- and post-intervention.[9]

METHOD

Participants

Fifty-seven patients (46 females and 11 males) were included in the corpus. They ranged in age from 20 to 82 years, with an average age of 53 ($\pm$ 14) years. Among the patients studied, five were diagnosed with Graves' disease (also known as Basedow's disease), six presented with thyroid cancer, 27 were identified as having multinodular goiter (GMHN), five were found to have hyperparathyroidism (HPT), two exhibited both multinodular goiter and HPT, and 12 were diagnosed with thyroid nodules (Figure 1). Forty patients underwent TT and 10 underwent partial thyroidectomy (PT). The remaining seven patients had either lymph node dissection or parathyroidectomy (removal of parathyroid glands), shown in Figure 2. In this study, participants were excluded if they failed to complete either the preoperative or postoperative voice recordings and the Voice Handicap Index (VHI) questionnaire,

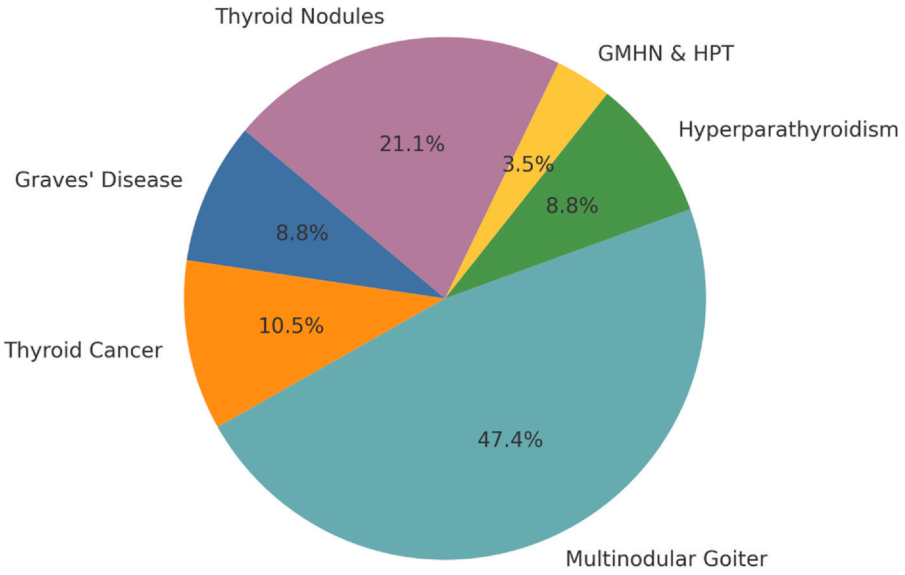

FIGURE 1. Diagnosis distribution among patients.

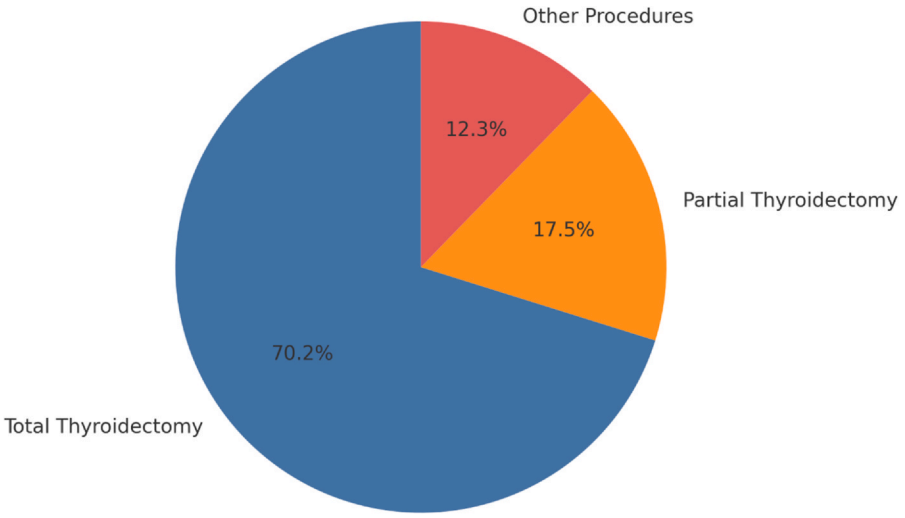

FIGURE 2. Surgical procedure distribution.

presented recordings of poor quality (e.g., recordings contaminated by loud background noise), or provided insufficient data due to either inability or unwillingness.

Operative and recording reports

The operative and recording reports were compiled for all patients by two surgeons involved in the study, author P.C. and Dr. Hélène Blaise from the University Hospital of Grenoble.

The data set thus includes information from throughout the surgery: patients' basics, surgery basics, recording log, subjective assessment. The surgical information comprises diagnostic details, surgical approaches (total or partial resection, parathyroidectomy resection), and intraoperative observations (nerve or muscle damage, as indicated by NIM neuromonitoring). Among the 57 patients evaluated, 54 experienced no nerve damage, with three reporting unilateral damage. Muscle damage reports were similarly skewed towards minimal occurrences: for 46 patients no damage was noted, for eight damage to one side of specific muscles (infrahyoid, cricothyroid, or omohyoid), and comprehensive muscle damage was observed in three cases. If the neuromonitoring is unchanged during the intervention, the nervous network is intact; the muscle damage is confirmed solely on the visual observation of the surgeon who may have damaged the infrahyoid muscle fibers (either dilacerated to remove a large goiter or invaded in the case of invasive carcinoma) or cricothyroid (most of the time by local invasion of a carcinoma).

During surgery, the vocal fold response to nerve stimulation is displayed on a nerve monitor as auditory or visual electromyographic (EMG) signals. Changes in the pattern of EMG signals, which can occur with retraction of the gland, dissection of surrounding tissues, or dissection of the RLN itself, can alert the surgeon to possible nerve irritation.[50,51]

Because vocal fold response is measured by EMG activities, only short-acting muscle relaxants, such as succinylcholine, should be used during induction of anesthesia when IONM is employed. No paralytic agents should be used after induction to prevent interference with obtaining an EMG signal.

Observations on morphological laryngeal pathologies were also recorded into the data set. The two main components of IONM are nerve stimulation and assessment of vocal fold response to nerve stimulation. The RLN can be stimulated with a low-voltage electric current delivered by a handheld probe (intermittent stimulation) or an electrode attached to the ipsilateral vagus nerve (intermittent or continuous stimulation). During planned TT, after completion of the initial lobectomy if IONM results suggest loss of function, the surgeon may consider stopping the operation for possible completion at a later date.[1] The laryngological examination was comprehensive for each patient. Out of 57 patients, 27 were reported no laryngological disorders, while the remaining 30 patients were diagnosed a variety of laryngological disorders. Among those 30 patients, 23 patients presented with minor troubles in laryngological mobility, two cases with complete paralysis, and five cases with various kinds of laryngeal lesions or abnormalities, in Figure 3.

The data set is divided into preoperative and postoperative records. All this information was collected by Sébastien Guigard from the original handwritten reports for each patient and collated in a spreadsheet database.

For this study, we focused on the information related to cervical endocrine surgery that was deemed relevant to the voice quality changes. This encompasses surgical categories, intraoperative injuries, postoperative recovery status, and patients' self-assessment.

Subjective examinations

Each patient completed a subjective assessment survey, the ten-question Voice Handicap Index (VHI-10).[52] The preoperative assessment was implemented the evening before the surgery. The postoperative survey took place 50 ± 21 days after the surgery. The VHI-10 is a widely recognized tool, considered relatively reliable by professionals in the field of voice.[53,54] Its purpose is to evaluate the impact of voice-related issues on the patient's perceived quality of life. It has been found that a difference of 6 on the VHI-10 may represent a minimal important difference, indicating the sensitivity of this index in capturing voice-related handicaps. Each patient read the evaluation individually before filling it out.

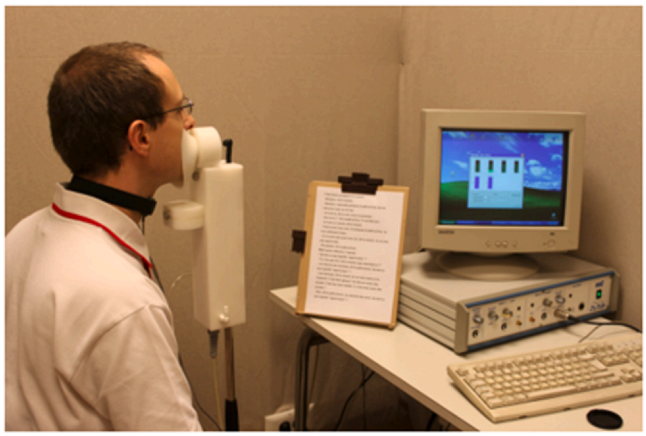

FIGURE 4. Depiction of the vocal recording setup equipment featuring the facial mask and EGG collar. EGG, electroglottographic.

Voice recording tasks

Data collection was conducted in a room of the hospital, treated with sound-absorbing walls (Figure 4). Patients were accompanied and instructed by the interns. EVA station (EVA 2, SQLab®, France)[55,56] was used to assess the aerodynamics and acoustics of speech and singing tasks. A facial mask, with a microphone and a flow meter, was affixed to the patient's face to capture audio signal, oral airflow signal. The voice sound level was calibrated in order to be able to compare data

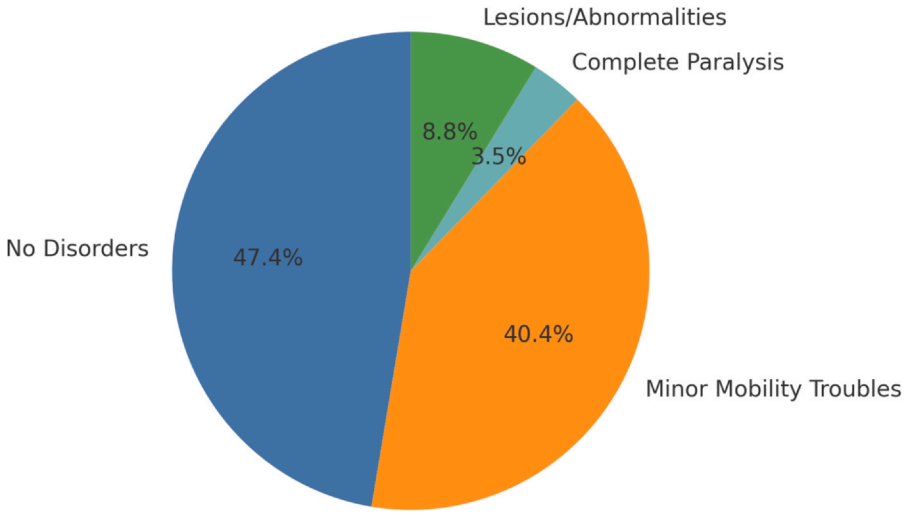

FIGURE 3. Laryngological disorders distribution among patients.

from one patient to another. The sound level at the standard AKG C1000 S microphone is calibrated as if it were 30 cm from the mouth. The calibration process was done by setting the source to 80 dB at 30 cm in an anechoic room using a B&K sound level meter, then adjusting the gain of the microphone. The aerodynamic measures are of little relevance to the voice quality metrics in this study, and will not be considered here.

Electrodes for measuring vocal-fold contact area by means of electroglottography (dual-channel EGG from Glottal Enterprises, www.glottal.com) were placed on the patient's neck at larynx height. All patients maintained a constant distance from the microphone and vocalized into a fixed mask to minimize measurement fluctuations.

The tasks considered in this study are as follows: text reading, sustained vowel /a/, glissando singing with sustained vowel /a/, and singing "Joyeux Anniversaire" ("Happy Birthday to You").

Data set processing

For this retrospective study, a subset of the existing recorded signals was first constructed. This subset consisted of the EGG signals and pressure-calibrated microphone signals. In the original data set, the audio was not level-calibrated but the audio level was available as a separate data track. Using this track, the microphone signal was calibrated for amplitude, and then upsampled. The audio signal was then combined with the EGG signal to generate a new two-channel audio file, for compatibility with the real-time voice-mapping analysis software FonaDyn version 3.[57]

The audio data were then batch-processed through scripts in FonaDyn, generating cycle-by-cycle log files and cell-by-cell (1 semitone × 1 dB) voice map files. The semitones are on the MIDI scale, where 0 ST corresponds to 8.1758 Hz and middle C (261.63 Hz) is 60 ST, according to the following formula:

$$f = 440 \times 2^{\left(\frac{m-69}{12}\right)} \quad (1)$$

where m is the number of the semitone on the MIDI scale.

FonaDyn detects phonation by measuring autocorrelation in the audio signal; its corresponding "Clarity" threshold was set to the default of 0.96 at a one-cycle delay, which is a fairly strict criterion for periodicity. Due to the high levels of noise present in the original EGG files, the EGG track was first denoised using a spectral thresholding function within FonaDyn set to about −40 dB relative to full scale in the EGG track. In the acoustic track, data with a calibrated audio SPL below 40 dB re 20 μPa were removed in order to suppress a persistent low-frequency hum.

The obtained voice map was further processed through MATLAB scripts for classification and visualization. The correlations between the data and patients' physiological and pathological information were computed in a spreadsheet program (Microsoft Excel).

Metrics and clustering

We submit that, rather than assessing each voice metric separately, the classification and subsequent mapping of combinations of metrics should facilitate a more clinically relevant qualitative analysis and interpretation of vocal function. FonaDyn version 3[58] incorporates a method for classifying phonation types by clustering combinations of several audio and EGG metrics.[9] For instance, very breathy voice is characterized by low spectrum balance (SB), low cepstral peak prominence (CPPs), high EGG cycle-rate sample entropy (CSE), an EGG contact quotient (Q_{ci}) close to 0.5 and a normalized peak EGG derivative (Q_{Δ}) close to 1. Here, we applied a previously described voice mapping and classification approach,[9] employing six acoustic and EGG metrics of general interest. The six metrics are the following (Table 1).

The *quotient of contact by integration* (Q_{ci}) is the area under the EGG pulse that has been normalized to 1 in both time and amplitude. The Q_{ci} was computed according to Ternström.[59] The Q_{ci} represents the relative amount of contacting during a cycle. A high Q_{ci} means that the vocal folds are in contact for a large fraction of each cycle. Prolonged contact indicates increased adduction and can suggest tension or hyperfunction in the vocal mechanism. Hyperfunctional voice disorders can result in vocal strain, fatigue, and potential voice damage over time. As such, an individual with a higher Q_{ci} after surgery might experience discomfort, fatigue, or even pain when speaking. This would likely lead to a higher perceived voice handicap, resulting in higher VHI scores. However, in soft or breathy phonation without VF contacting, the EGG waveform is of low amplitude and practically sinusoidal, causing Q_{ci} to

TABLE 1.
The EGG and acoustic metrics chosenAcoustic Metrics Chosen as Input Features, and the Ranges That Were Mapped to the Interval [0…1] Prior to Clustering

Type	Symbol	Definition	Range
EGG metrics	Q_{ci}	Quotient of contact by integration	0-1
	Q_{Δ}	Normalized peak derivative	1-10
	CSE	Cycle-rate sample entropy	0-10
Acoustic metrics	Crest Factor	Ratio of the peak amplitude to the RMS amplitude	3-12 dB
	SB	Spectrum balance	− 40 to 0 dB
	CPPs	Cepstral peak prominence smoothed	0-15 dB

become close to 0.5. In this case, Q_Δ needs to be taken into consideration together with Q_{ci} in order to find out the contacting status.

The *normalized peak derivative* (Q_Δ) is the maximum derivative over each EGG cycle.[59] It represents the maximum rate of contacting during closure. Q_Δ was computed as follows:

$$Q_\Delta \approx 2\delta_{\max} / \left[A_{p-p} \cdot \sin\left(\frac{2\pi}{T}\right) \right] \tag{2}$$

where T is the period length in integer samples, $A_{p\text{-}p}$ is the peak-to-peak amplitude, and δ_{max} is the largest positive difference observed over the period between two consecutive sample points in the discretized EGG signal. In phonation without vocal fold collision, the EGG becomes a low-amplitude sine wave. Hence the minimum value that Q_Δ can assume is the peak derivative of a normalized sine wave, which is 1. The Q_Δ represents the maximum rate of contacting during closure of the vocal folds. Low values suggest incomplete or less efficient vocal fold closure, which could result in a breathy or weak voice. Inefficient vocal fold closure can impact voice quality and vocal stamina, potentially leading to higher VHI scores.

The cycle-rate sample entropy (CSE)[60] is a perturbation metric.[61] CSE is low for a regular, self-similar signal and high when a signal is transient, erratic, or noisy. CSE represents the cycle-to-cycle instability of the EGG waveform. The CSE was computed over a short sliding window of glottal cycles, on the basis of the four first Fourier descriptors (four levels and four phases) of the EGG pulse waveform, with a scaling from 0 (very regular) to 10 (very disordered). Higher CSE values indicate increased variability in EGG pulse shape. CSE is invariably high in breathy (noncontacting) phonation, also in healthy voices. Greater irregularity in vocal fold vibration in the presence of contacting could imply a variety of voice issues related to incomplete or overly forced closure of the vocal folds. Perturbations are known to result in a voice that sounds rough, hoarse, or unstable, leading to challenges in daily voice use and communication. Therefore, an individual with a persistently high CSE might perceive a more significant handicap in their voice, resulting in higher VHI scores.

The audio *crest factor* is computed as the ratio of the peak amplitude of the RMS amplitude for every phonatory cycle. It is a simple indicator of the peakiness of the voice signal and tends to be especially high in creaky voice. A low Crest factor in the vocal range of habitual speech indicates indirectly a less distinct interruption of glottal flow at closure. Such a voice might be perceived as being less clear or of lower quality, which could lead to higher VHI scores.

The spectrum balance (SB) is here defined as the difference in acoustic power level (dB) above 2 kHz and below 1.5 kHz.

$$SB = 10 \cdot \log_{10}\left(\frac{W_{>2kHz}}{W_{<1.5kHz}}\right)(dB) \tag{3}$$

The SB is typically < 0 and increases (becomes less negative) when the relative amount of high-frequency energy in the signal increases. It is indirectly related to the maximum second derivative of glottal flow.[62] The SB can be affected also by vowel articulation, and, at low signal levels, by system noise in the audio chain. A low (very negative) SB indicates that there is less energy in the higher frequencies relative to the lower frequencies. This could be indicative of a less clear, duller voice when high frequencies contribute to voice clarity and intelligibility. Damage to the vocal folds (indirectly mostly), arytenoid muscles, and associated nerves might lead to a reduction in the SB and thus lead to a higher VHI scores.

The *cepstral peak prominence smoothed* (CPPs)[63] is a measure of periodicity in the acoustic *spectrum*. The higher the CPPs, the stronger the harmonics and the weaker the noise in the audio signal. Here, the calculation of smoothed CPPs followed Awan et al[64] (7 frames in time, 11 bins in quefrency). A lower value of CPPs indicates more noise and weaker periodicity. A voice that lacks harmonics or has added noise can be perceived as less stable or clear, which might lead to a higher VHI score.

All these metrics tend to increase in value as voicing becomes louder, more regular and/or less pathological; except CSE, which tends to decrease.

Some of these metrics are evaluated cycle-synchronously: Q_{ci}, Q_Δ, CSE, crest factor, and SPL. Others are windowed with a fixed frame length: CPP (23 ms), SB (fourth-order low-pass smoothing at 50 Hz), and f_o (23 ms). FonaDyn stores observations of the six metrics once for every phonated cycle. These observations are then averaged cell by cell in the voice map, which has a separate layer for each metric. Hence the number of observations contributing to each voice map is on the order of 10^5.

For clustering, we applied the K-means++ algorithm following the routine described by Arthur and Vassilvitskii.[65] Metric values were standardized to the range [0…1] before clustering. Q_Δ only was first transformed to its base-10 logarithm, because of its skewed distribution. We choose an initial center c_1 at random from the dataset, then selecting the next centroid with probability:

$$\frac{d^2(x_m, c_1)}{\sum_{j=1}^{n} d^2(x_j, c_1)} \tag{4}$$

where the distance between c_1 and the observation m is denoted as $d(x_m, c_1)$. Then the Euclidean distances from each new observation to each centroid are computed, and the nearest centroid location is updated until the algorithm has reached a stable state over iterations.

To this end, we apply the K-means-based classification method to the dataset. We aggregate all preoperative and postoperative recordings for each patient into one contiguous recording, and then utilize clustering algorithms to automatically identify the centroids in a six-dimensional metrics space. The reason we calculate the difference of metrics pre- and post-operation is because every patient has his/her own voice features before and after operation. By focusing on the difference or change, we can control for

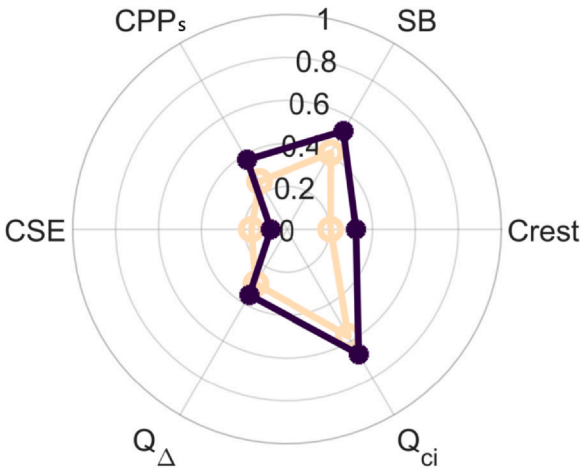

FIGURE 5. Radar plot of centroid values of a two-cluster classification.

this individual variability and can measure the impact of the surgery on voice parameters. This difference provides a clearer picture of the surgical effect, separate from any baseline tendencies. Constraining the comparison to the voice range that overlaps pre and post removes the bias that would be likely from changes in habitual f_o and/or SPL.

Each such centroid then represents a distinct mode of phonation. The occurrence of each centroid is represented as a 2D distribution over f_o and SPL. If the type of phonation changes, the distributions will also change. Subsequently, we will map these centroids back onto the preoperative and postoperative conditions to observe any variations that occur relative to the cluster centroids, postsurgery.

A centroid can be conveniently graphed as a radar plot (Figure 5). Each radial spoke of the radar plot serves as the axis of a standardized metric. The values on these axes are scaled between 0 and 1, indicative of the proportional representation of each metric. This can also be interpreted as 0% to 100%.

Voice map-based comparisons

In the conventional VRP, the attainable extremes in f_o and SPL are dependent variables (dependent on vocal status), while when making voice maps, f_o and SPL can be seen as *in*dependent variables on which any given metric is dependent. When we observe a person's voice map of individual or combined metrics, we invariably see distinct subregions formed on the two axes of f_o and SPL. Referring to the Figure 3 of Cai et al,[9] five subranges across the SPL range can be discerned, based on combinations and trends of the voice metrics. For example, subrange A is characterized as high CSE, low crest factor, CPPs, and Q_Δ, and a Q_{ci} of 50%, which is a typical sign of breathy voice by definition. We can thus name subrange A as "breathy." Similarly, the subranges B-E could be referred to as: "transition," "loose," "firm," "hard." By grouping the phonation types based on combinations of metrics, salient features are more easily identified. Notice that individuals could exhibit other patterns than in the case above, and for pathological voices, the manifestations can be even more complex.

Both f_o and SPL exert a large bias on all these metrics[49]. The changes of voice metrics correlate strongly to changes in f_o and SPL. To assess accurately any changes in these metrics it is therefore necessary to make the pre-post comparisons at matched f_o and SPL. This can be conveniently done by creating *difference* maps of each metric. Figure 6 shows an example of two overlapping regions from a pre- and a post-operative recording. Differences can be computed only within the intersection of the two regions.

Here each point represents 1 semitone × 1 dB cell; the "post" is marked as "+" and the "pre" is marked as "T." The gray rectangular cells make up the overlapped region that occurs both in the "pre" and the "post." This allows the calculation of changes in various aspects of voice quality throughout the overlap area, that is, in that part of the voice range that could be elicited both pre- and post-intervention. As mentioned above, such a comparison then accounts for the inherent and person-specific dependencies of the given metric on f_o and SPL.

Unless otherwise noted, the following analyses were done by mapping changes in the pre-post overlap area only, and are thus as independent as possible of the changes in the voice range contour. The size of any area on a voice map is conventionally given in cells, or [ST × dB].

Statistics

Data underwent preliminary checks for missing values and distribution normality.

Descriptive statistics were computed for all variables. The difference is typically calculated as the postcondition minus the precondition. The 95% confidence interval for the difference between the two conditions is:

$$CI = (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,df}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (5)$$

where $\bar{x}_1$, $\bar{x}_2$ are the sample means of the two groups. s_1^2 and s_2^2 are the sample variances, n_1 and n_2 are the sample sizes. $t_{\alpha/2,df}$ is the critical t-value from the t-distribution corresponding to $\alpha/2$ and the degrees of freedom df.

Correlations between continuous variables were assessed using Pearson's coefficients.

The optimal number for phonation clusters is chosen by the Bayesian Information Criterion (BIC). It measures the model fit while penalizing models with more clusters. The BIC is formulated as:

$$BIC = -2\ln(L) + k\ln(n) \quad (6)$$

where ln is the natural logarithm, L is the maximized value of the likelihood function of the model, k is the number of clusters, and n is the number of observations.

All statistical analyses were conducted using Microsoft Excel and the Python SciPy package.

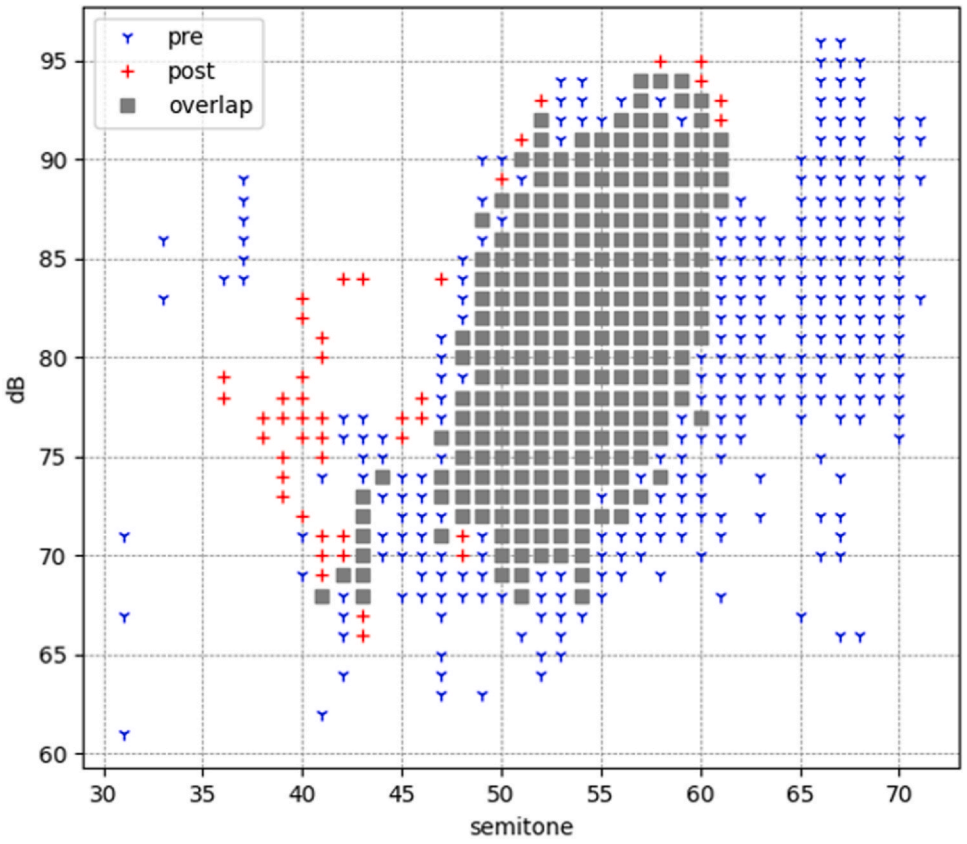

FIGURE 6. Example of overlapping voice maps, pre-post intervention. The horizontal axis is f_o in semitones on the MIDI scale; the vertical axis is unweighted sound level at 0.3 m in dB re 20 μPa.

RESULTS

Range, quality metrics, and clusters

Voice range

The predominant effect observed post-thyroidectomy is a reduction of the voice range. As expected, the results in this study showed significant changes in the habitual and attainable ranges of f_o and SPL.

From Table 2 we clearly observe that post-thyroidectomy, there is a general decline in f_o and SPL, with increased variability. All the aspects of the voice range extremes also tend to decrease. This result aligns with findings from other studies that the voice capabilities may deteriorate following the surgery.

While the majority of effects were small, outliers and deviation in the aggregated data suggest that some individuals experienced substantial changes in their voice range profiles. The distribution of voice range changes can also be seen in Figure 7. Although a reduction of voice range was the most common case, quite a few patients actually had a larger voice range after the intervention.

TABLE 2.
Voice Range-related Changes, Pre- and Post-operation

Parameter	Prevalue (Mean ± Std)	Postvalue (Mean ± Std)	Diff-mean	*P*-value	95% CI
Average f_o (ST)	56.56 ± 4.32	53.81 ± 5.55	− 2.75	***	[− 4.03, − 1.47]
Average SPL (dB)	86.13 ± 3.28	84.64 ± 3.94	− 1.49	*	[− 2.42, − 0.56]
f_o min (ST)	36.23 ± 5.21	34.40 ± 5.17	− 1.79	*	[− 3.46, − 0.13]
f_o max (ST)	72.09 ± 5.42	68.32 ± 8.72	− 3.65	**	[− 6.47, − 0.84]
SPL min (dB)	62.65 ± 3.19	62.21 ± 2.56	− 0.44	0.36	[− 2.57, 1.69]
SPL max (dB)	103.54 ± 4.26	102.07 ± 5.13	− 1.43	0.06	[− 4.93, 2.07]

***Highly significant ($P < 0.001$).
**Very significant ($P < 0.01$).
*Significant ($P < 0.05$).
No asterisk: not significant.
SPL, sound pressure level.

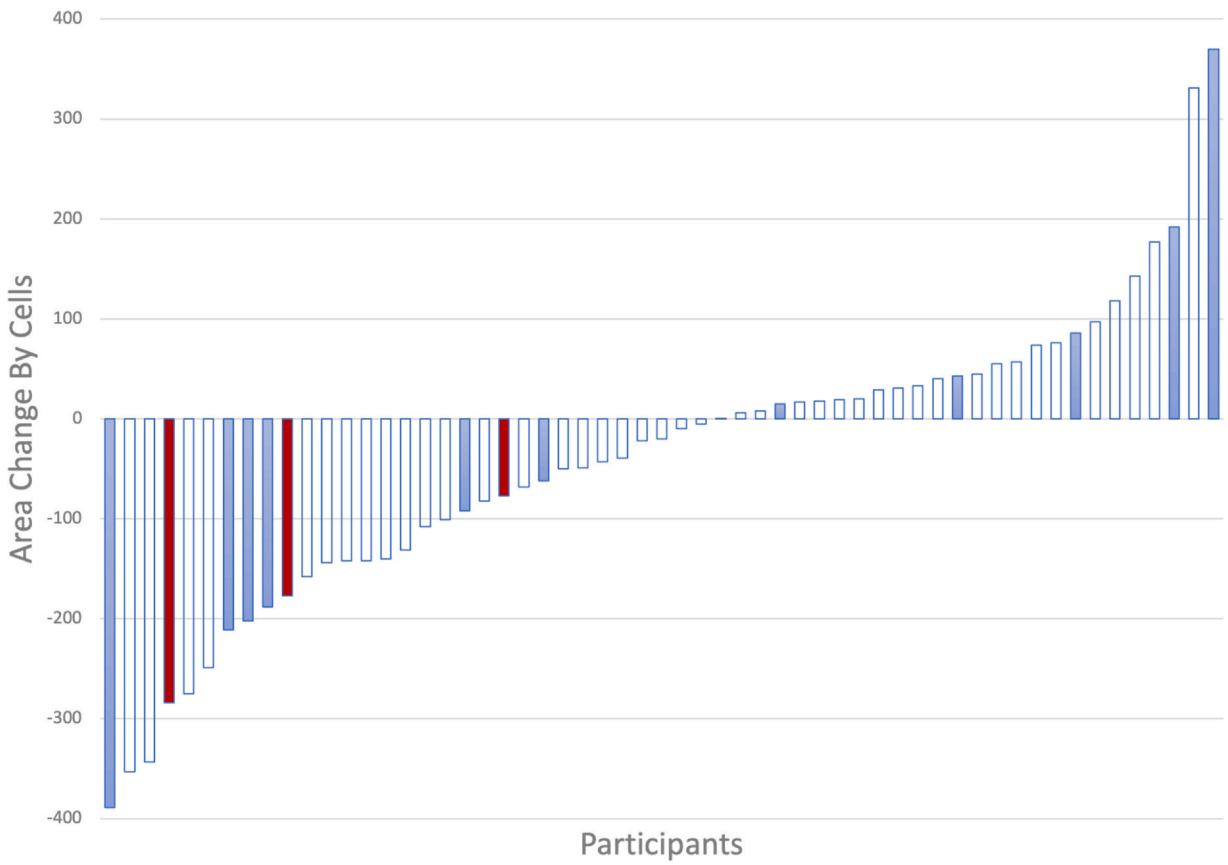

FIGURE 7. Distribution of voice range change by participant. Blank bars represent patients without damage in nerve or muscle. Blue bars represent patients with damage in muscles. Red bars represent patients with damage in nerves. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Voice metrics

The descriptive statistics of the six metrics pre- and post-thyroidectomy, taken across all participants and over their entire overlap areas, are given in Table 3.

A significant decrease was observed in SB ($P < 0.05$), CPPs ($P < 0.05$), and Q_Δ ($P < 0.001$), which, while consistent across the population, was numerically small. An increase was noted for CSE postoperatively ($P < 0.01$), yet probably too small to be clinically relevant. The distribution plots in Figure 8 also show very little change, with only a smaller range of extremes in the metrics CSE and Q_Δ.

These changes appear to be minimal, with negligible alterations in the mean, having little impact at the group level on acoustic perception and vocal functionality. However, the population statistics may be obscuring individual variations. Each individual's intraoperative condition varies, so aggregating patients who have recovered well with those who have not into a single set may cause the effects in individuals to cancel each other out. However, when examining the impact of the surgery itself on patients, these significant differences shown above still offer insights into potential postoperative trajectories. In any case, they

TABLE 3.
Metric Differences in Pre- and Post-conditions

Name	Premean	Postmean	Diff-mean	*P*-value	95% CI
Crest Factor	2.10 ± 0.17	2.08 ± 0.19	− 0.01	0.37	[− 0.003, 0.008]
SB (dB)	− 27.76 ± 3.53	− 28.04 ± 3.07	− 0.27	*	[− 0.33, − 0.11]
CPPs (dB)	8.71 ± 2.11	8.53 ± 2.18	− 0.17	*	[− 0.21, − 0.1]
CSE	3.18 ± 1.28	3.27 ± 1.30	0.09	**	[0.04, 0.09]
Q_Δ	6.37 ± 1.16	6.08 ± 1.04	− 0.29	***	[− 0.38, − 0.21]
Qci	0.43 ± 0.04	0.43 ± 0.03	0.00	0.75	[− 0.00, 0.002]

***Highly significant ($P < 0.001$).
**Very significant ($P < 0.01$).
* Significant ($P < 0.05$).
No asterisk: not significant.

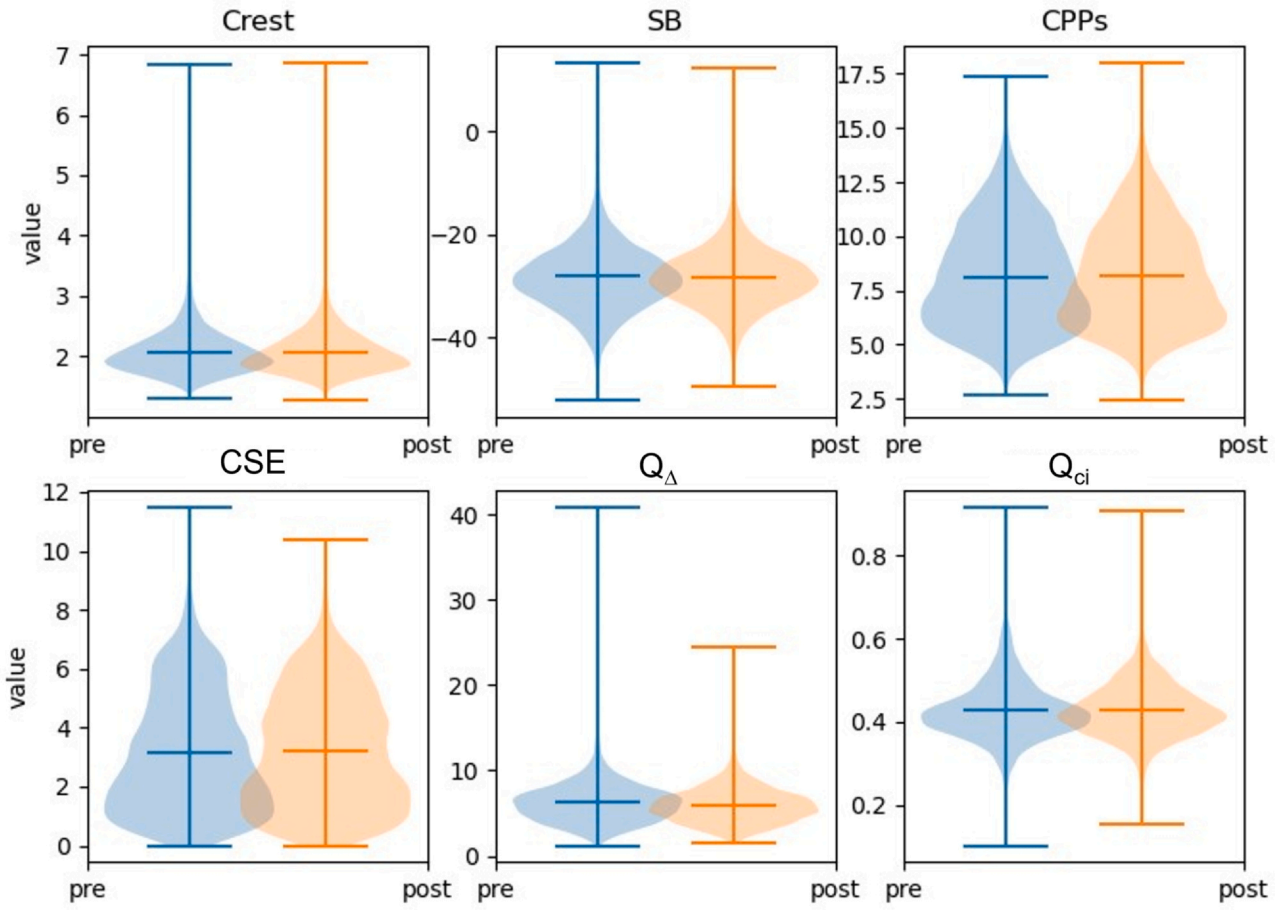

FIGURE 8. Violin plots for the individual voice metrics, pre- and post-intervention.

represent the benefits of controlling the variables for f_o and SPL.

This lack of pre-post differences at the group level does not imply that individual patients were not affected. Also, each metric was here taken in isolation.

Phonation type cluster change

The six metrics have a certain degree of covariation; and not all combinations of values are physiologically possible. Using statistical clustering, we can identify the most prevalent combinations. The centroid of each cluster is the vector of six metric means that characterizes the combination. By assigning a unique color to each cluster, and plotting those colors across f_o and SPL, the behavior of the voice under study can be documented.

We observed that for the majority of patients, a solution with two clusters suffices to exhibit clear differences. Figure 9 shows an example of how the clusters trained on one patient manifested in a voice map.

Here the colored regions are overlapping areas that occur in both pre- and post-conditions for this patient. Each grid cell represents the occurrences of a matched f_o and SPL. The actual voice range is of course larger than that represented here in the illustration. Of these two clusters, one is characterized by higher Crest Factor, higher CPPs, higher SB as the acoustic features, and lower CSE, higher Q_Δ, slightly higher Q_{ci} as the EGG features. Received knowledge lets us define this cluster as "stability." This combination is what we found to be the most frequent one. Conversely, the other cluster was labeled as "instability."

The overlap region before the surgery is dominated by the "stability" cluster (marked as dark blue, 2), while after the surgery the overlap region is dominated by the "instability cluster" (marked as pale yellow, 1). This indicates that postoperative voice quality tends to become worse and less stable.

By counting the cells that each cluster receives, we obtain a general quantification of the degree of change across the operation. The centroid values for each patient may vary, yet they exhibit similar patterns, as demonstrated in Figure 9. This consistency allows for the cluster regions to be comparably analyzed across subjects.

In "post," the "stability" cluster region was 6% smaller than in "pre" (mean of 57 patients). The centroid positions also changed. On average, the f_o of the stable cluster centroid dropped by 1.05 semitone and the SPL increased by 0.86 dB.

Above we discussed the two-clusters outcome. If we increase the number of clusters to three, then for some of the patients we observe different patterns.

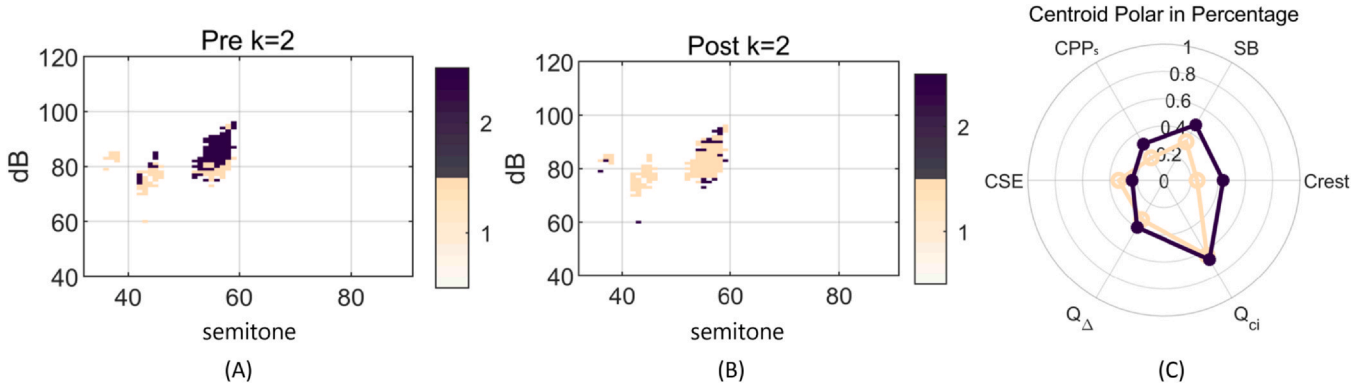

FIGURE 9. Assessing voice change by mapping clustered metrics. Only the pre-post overlap region is shown. Example for one female patient who had total thyroidectomy and worse voice quality after the operation. (**A**) Preoperative, (**B**) postoperative, (**C**) radar plot for the centroids obtained by combined clustering of the signal metrics from all productions, pre and post. Note how cluster 2 (dark blue) dominates preoperatively and cluster 1 (beige) dominates postoperatively. Examining the metrics shows that cluster 2 represents a voice that was more stable and richer in high frequencies than that for cluster 1 (see text). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

In this patient's voice map (Figure 10), still the dark blue (3) is the relatively stable region characterized as higher Crest Factor, higher CPPs and lower entropy. The beige (1) is the unstable part characterized by higher CSE, high SB; while the pale blue region, (2) has lower SB and smaller Crest Factor than the unstable cluster. Here we refer to the latter as a "transition" region, representing an improved status compared to the unstable region.

In the preintervention condition, this voice map of clusters is divided into three regions with the relatively unstable region in color beige appearing at the lower voice range. In the postoperative comparison voice map, we note that the "unstable" region has almost disappeared and been replaced by the "transition" region. The voice map is now covered by the "transition" and "stable" regions. Notably, the stable region is also expanded, occupying half of the voice map postsurgery. This agrees with the auditory impression from the actual recordings of less breathiness (though it does not mean that this patient can do better in all of the tasks in this study; she lost her high pitches in singing and found it hard transitioning between registers).

Voice Handicap Index

The Voice Handicap Index (VHI) is a self-reported subjective measure of voice-related difficulties in daily life. In this section, we examine whether there is a correlation between this subjective scale and the objective measurements.

VHI changes and voice range

We first analyze the VHI scores preoperatively and postoperatively. Out of the 57 patients studied, 25 exhibited a deterioration (increase) in their VHI score. Conversely, six patients demonstrated an improvement (decrease) in their VHI score. The remaining 26 patients showed no change in their VHI scores. Notably, among these 26 patients with unchanged scores, 10 reported no voice issues both preoperatively and postoperatively (VHI = 0). Preoperatively, the mean VHI score was 3.12 with a standard deviation of 4.26. The range of scores spans from 0 to 19. Postoperatively, the mean VHI score was 7.63 with an increased standard deviation of 7.34. The range of postcondition ranges from 0 to a notably higher maximum value of 31. It is evident that thyroidectomy has a

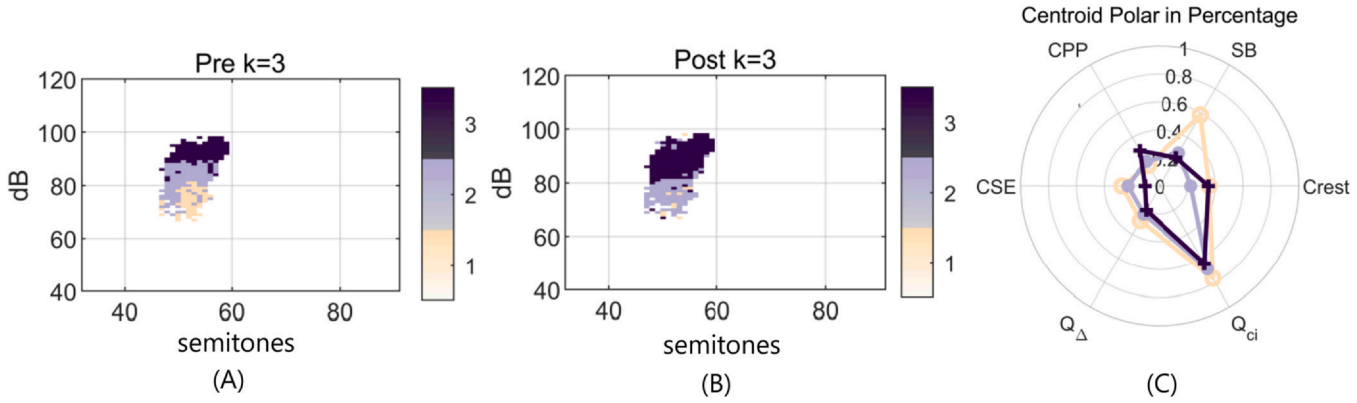

FIGURE 10. Patient No. 34. Clustering on voice maps, $k = 3$. (**A**) preoperative, (**B**) postoperative, (**C**) radar plot for the centroids. Note how cluster 1 (beige) decreases in size postoperatively. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

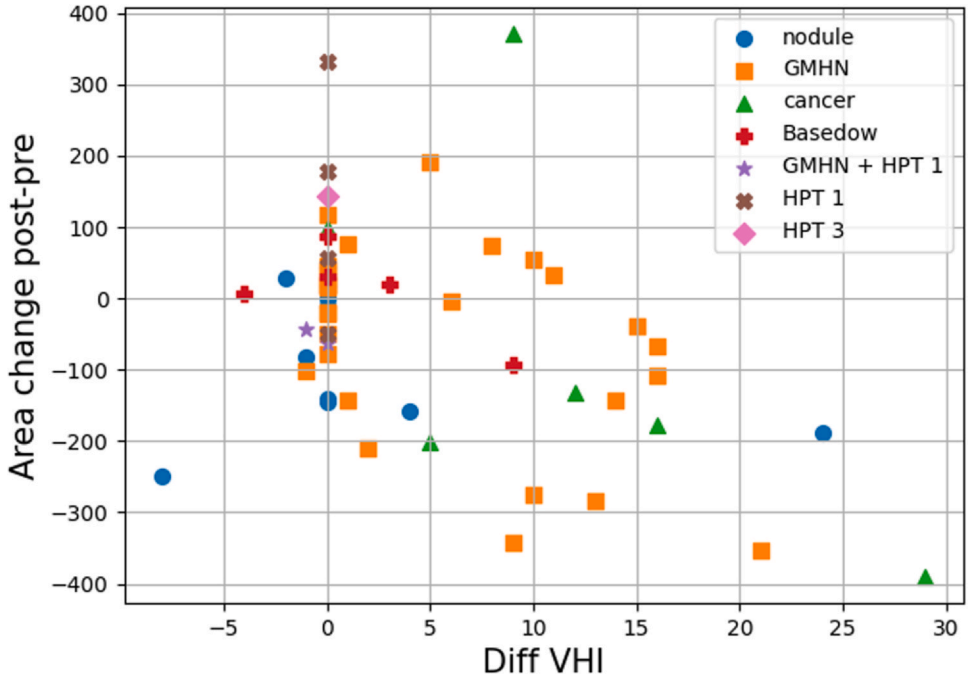

FIGURE 11. Scatter plot of VHI changes versus voice range change by diagnosis. VHI, Voice Handicap Index.

worsening effect on the VHI scores (P-value = 0.01). The widened score range in the postoperative phase indicates a heterogeneity in patient responses following the surgical procedure.

Thyroidectomy typically worsens VHI scores, indicating a decline in voice quality for many patients, which is in agreement with earlier studies.

The voice complaints as represented by the VHI-10 correlated primarily with a reduction of the voice range. There was a moderate negative correlation between changes in the VHI and voice range; specifically, a reduced voice range corresponded to a heightened VHI score. The Pearson correlation coefficient is -0.44 ($P < 0.05$). Every 40 cells (ST × dB) decrease in postoperation voice range results in an increase by 1 in VHI (Figure 11).

Anecdotally, patients as well as clinicians often recognize the loss of higher or lower pitches more readily than changes in voice quality. Patients tend to express greater concern regarding the reduction in voice range than with changes in voice quality. The VHI is closely linked to changes in the voice range area. One explanation for this is that while patients have no immediate voice concerns about alterations in voice quality, they are more likely to notice challenges when attempting to pronounce at specific SPL and f_o. This difficulty will manifest as a decrease in the area represented in the VRP.

In addition, patients with different diagnoses exhibit distinct patterns of change in VHI and voice range. For instance, the cancer and GMHN groups show trend of increased VHI differences with decreased voice range changes. Conversely, the Basedow group's minimal voice range changes suggest a lesser impact on their perceived voice range.

VHI changes and the voice metrics

In the analysis of voice metrics, we observed no correlation between VHI scores and any individual metric.

We can observe in Figure 12 some weak linear correlations: lower values of Crest Factor, SB, CPPs, and Q_Δ were associated with higher VHI differences. Conversely, higher values of CSE and Q_{ci} correlated with higher VHI differences.

In Table 4, none of these metrics were significantly correlated with the VHI ($P > 0.05$), suggesting that while there might be observable trends that are in accordance with what would be expected, they are not robust enough to be conclusive. Furthermore, since the VHI is a self-assessment scale and the mentioned metrics are objective, there might be discrepancies. A patient might obtain a low score due to personal dissatisfaction, even if their metrics are improved, or vice versa.

VHI changes and cluster changes

While a solitary metric might provide some insight into the trend of voice complaints, the combined trends in several metrics may have greater explanatory power. A single metric might not adequately describe the trend of voice complaints. However, if the metric meets the constraints of other metrics, they might align with voice complaints and also pertain the interpretability. For instance, a Q_{ci} value of 0.5 is ambiguous, in that it can represent both more contacting than normal, and no contacting at all. This is easily resolved by combining it with Q_Δ, which is close to 1 for no contacting and >2 when there is some contacting. Still, both are needed: Q_Δ does not inform on the degree of contacting, only on how fast it changes over the phonatory

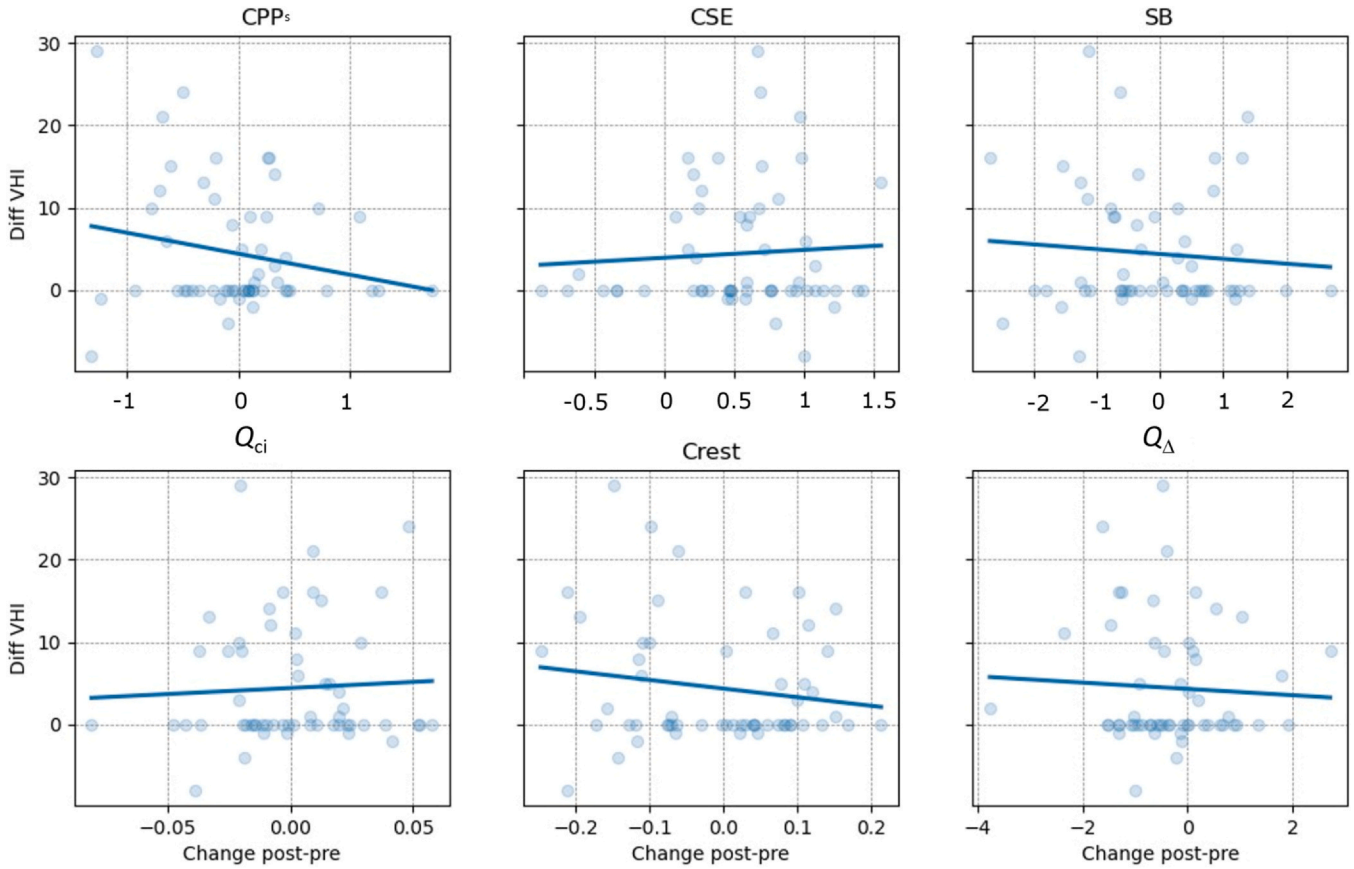

FIGURE 12. Scatter plots and linear regression of the VHI versus the voice metrics. From the top-left to the bottom-right, they are respectively CPP, CSE, SB, Q_{ci}, Crest, and Q_{Δ}. CPP, cepstral peak prominence; CSE, cycle-rate sample entropy; SB, spectrum balance; VHI, Voice Handicap Index.

TABLE 4.
Correlation Between Metrics and VHI Differences

Name	Diff-VHI correlation coefficient	*P*-value	95% CI
Crest Factor	− 0.21	0.11	[− 0.04, 0.02]
SB (dB)	− 0.08	0.54	[− 0.99, 0.47]
CPPs (dB)	− 0.23	0.09	[− 0.34, 0.29]
CSE	0.09	0.50	[− 0.19, 0.36]
Q_{Δ}	− 0.08	0.53	[− 0.61, − 0.05]
Qci	0.07	0.62	[− 0.01, 0.01]

No asterisk: not significant.

cycle. Combinations/clustering of the metrics across overlapped cells showed some weak correlation with the VHI score, with a coefficient of − 0.2 ($P = 0.05$). In other words, the more the voice remains within the "stability" cluster postoperation, the higher the likelihood of patients reporting improved voice quality.

Comparatively, the "stable cluster area" was more correlated to voice complaints than was any individual voice quality metric; although still less significant than the correlation seen with changes in the voice range.

The findings underscore the importance of a multimetric evaluation in understanding voice complaints. Relying solely on individual metrics and the average outcome of all the tasks may lead to a fragmented view, possibly overlooking some nuanced aspects of voice quality. The clustering of metrics offers a more integrated perspective. Even though the coefficient is modest, the indication that stability in voice (as defined by the change in stable cluster relative area) contributes positively to the patient's postoperative experience is valuable. This insight paves the way for further exploration into how surgical interventions can aim for such stable zones to enhance postoperative voice quality satisfaction.

Case studies

We now consider two patient cases. The first, a 52-year-old female, underwent a TT, supposedly without nerve or muscle damage. Presurgery, she had no issues performing the tasks, reflected by a VHI score of zero. However, 38 days postsurgery, she reported voice alterations, primarily a loss of high-pitched voice quality and reported vocal fatigue, leading to a 10-point increase in VHI. As per Figure 13, there was a reduction in voice range (by 275 cells), concurring with the VHI increase.

Figure 14 shows the changes in the acoustic and EGG metric differences. A higher SB (+ 0.69), a lower CPPs (− 1.56), a higher CSE (+ 0.35), a higher Q_{ci} (+ 0.03) were

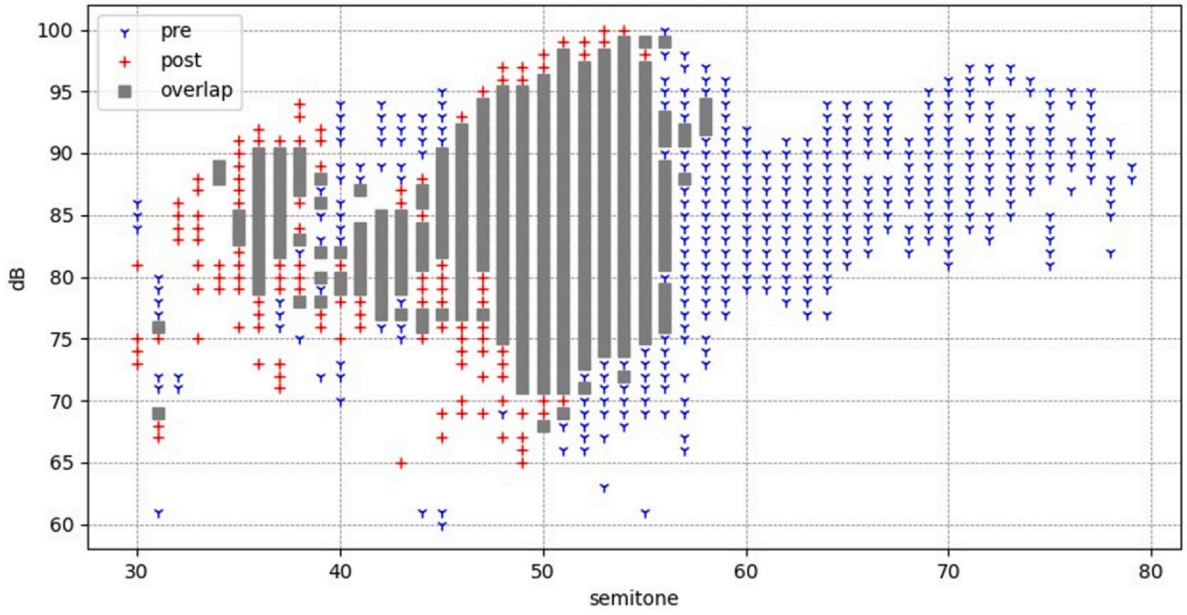

FIGURE 13. Voice maps pre and post for Patient No. 6. "Pre" voice map is marked in blue, and "post" voice map is marked in red crossing. The rectangular gray cells are the overlapping area. Examine how higher pitches decrease postoperatively. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

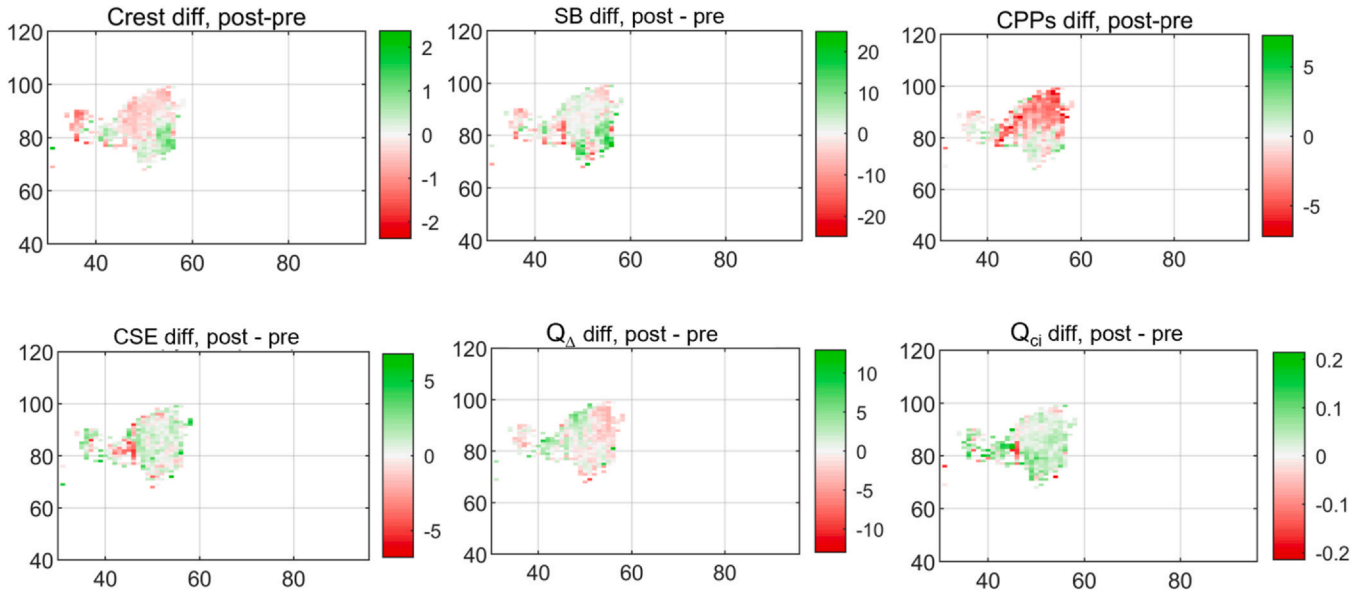

FIGURE 14. Patient No. 6. Acoustic and EGG metrics differences (post minus pre), the green color means increasing and the red color means decreasing. The horizontal axis is fo in semitones, the vertical axis is SPL in dB @ 0.3 m. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.). SPL, sound pressure level.

observed after surgery, suggesting voice instability. The values here are calculated by the mean of all cells in the overlap area. At high SPL and high f_o, the CPPs decreased mostly, which reflects the patient's challenges in phonating loudly and highly after the surgery.

Figure 15, more specifically, shows that at specifically low SPLs, the Q_{ci} shows a change from green to yellow, indicating an increase from around 0.4 to 0.5 post-operation. When the vocal folds are vibrating softly and without actual colliding, the EGG signal is very weak and almost sinusoidal, which results in a Q_{ci} of 0.5. This again indicates a worsened voice quality.

In Figure 16, we also find that the clustered metrics give an integrated aspect of the deteriorated voice in terms of the decreased relative area of the "stable" region (in dark blue). Also, on the postvoice map, the top right part that represents the high SPL and f_o transitions from the "stable" region to "instable" one.

There were also cases that showed contradicting results in terms of voice metrics and VHI. A 49-year-old female

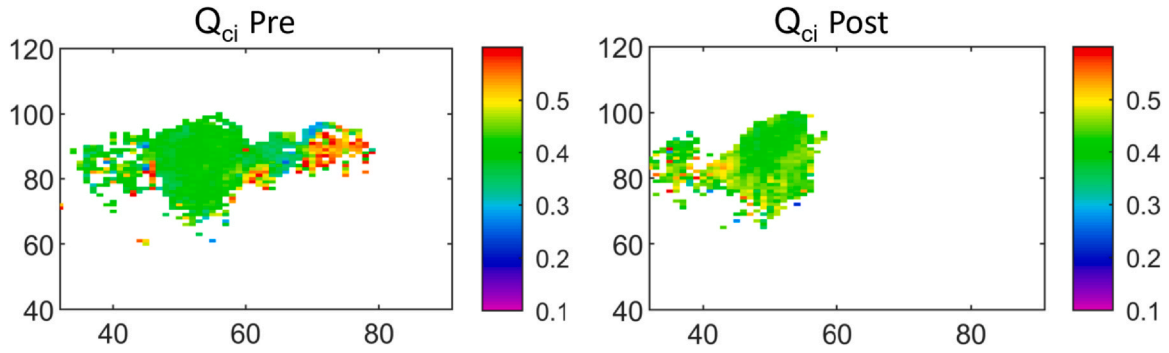

FIGURE 15. Patient No. 6. Q_{ci} voice maps pre and post. The average of Qci across all cells increases by 0.03 postoperatively. The horizontal axis is fo in semitones, the vertical axis is SPL in dB @ 0.3 m. SPL, sound pressure level.

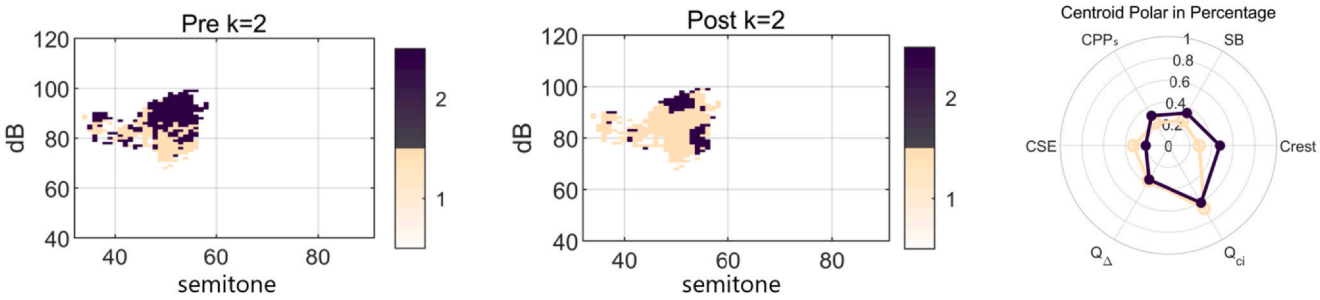

FIGURE 16. Patient No. 6. Voice maps of cluster regions pre and post and the corresponding metrics radar plot of the two cluster centroids.

patient had TT without reported nerve and muscle damage. In the prerecording, the patient showed voice struggle as indicated by a VHI score of 16. Though facing persistent difficulties in some of the tasks and glissandi, she felt herself as having an improved voice after the surgery, being less hoarse postsurgery, with an improved VHI score of 8. More specifically, certain questions show marked improvement as F1 decreased from 3 to 2, P10 from 3 to 1, P14 from 3 to 1, P17 from 4 to 1 (see Appendix A). Contradictorily, the acoustic measurements are not consistent with this. The voice range decreased by 249 cells.

When observing the metric difference map in Figure 17, the voice metrics all indicate a detrimental change: all metrics decrease postsurgery, except CSE for which the polarity is the opposite. All these changes typically mean less efficient or forceful closure, decreased voice quality after the surgery.

Figure 18 further shows the decline in the "stable" voice area (in dark blue), corroborated audibly as a breathier,

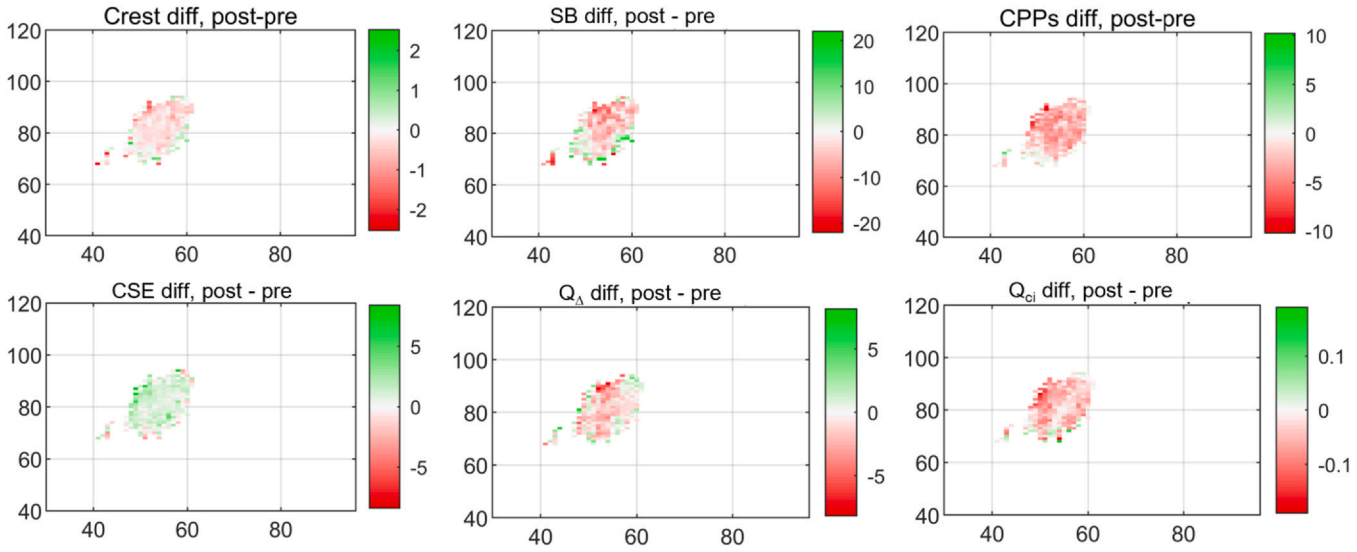

FIGURE 17. Patient No. 1. Acoustic and EGG metrics differences (post minus pre), the green color means increasing and the red color means decreasing. The horizontal axis is fo in semitones, the vertical axis is SPL in dB @ 0.3 m. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.). SPL, sound pressure level.

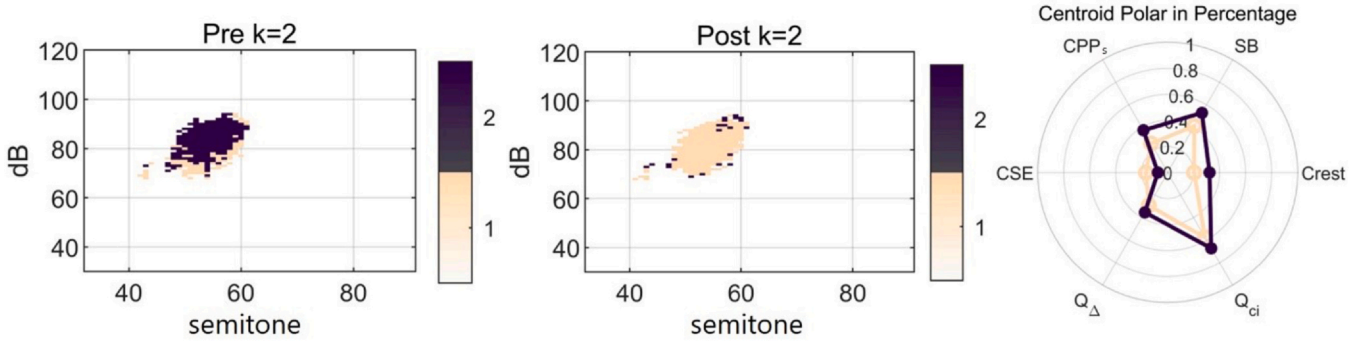

FIGURE 18. Patient No. 1. Voice maps of cluster regions pre and post and the corresponding metrics radar plot of the two cluster centroids.

fatigued voice postsurgery. The patient due to psychological issues was afraid of overusing her voice after the operation. Additionally, a papillary thyroid cancer and subsequent iodine treatment left her emotionally strained during the second recording.

Surgery types and groupings

We now partition the dataset by surgical indication and by quantitatively indicated outcomes.

Partial versus TT

Table 5 shows a breakdown of the results by type of surgery, that is, TT or PT. The lymph node dissection as a type was discarded because there were too few instances of it in the data set.

As shown in Table 5, compared to the TT group, the PT group exhibited a lower increase in VHI, indicating fewer voice complaints or better-perceived voice quality, and a smaller voice range decrease. These differences were not significant ($P > 0.05$). However, the 95% CI for VHI difference shows a clear decrease in PT comparing to TT groups.

Among the voice metrics, SB shows decrease ($P < 0.05$) in the PT group and increase in the TT group. The CSE increases ($P < 0.05$) in the PT group and slightly decrease in the TT group. While in the TT group the Q_{ci} stays unchanged, in the PT group the Q_{ci} increased a little ($P < 0.05$). Those metrics show significant difference between the PT and TT groups. However, these differences are marginal.

Apart from these, the other metrics show no statistical significance. The crest factor, CPPs, and Q_{Δ} decrease in PT, indicating potential worsening in PT group, though not statistically significant.

"Improved" and "deteriorated" group

Drawing from the observations in Section "VHI changes and cluster changes" regarding cluster changes, we categorized subjects based on their postoperative vocal outcomes as either improved or deteriorated. Essentially, when a cluster associated with higher voice quality (like higher CPPs and lower CSE) postoperatively occupies a greater area in the overlap region, we interpret this as an improvement in voice and classify such subjects under the "improved" group. Patients whose "stable" cluster area shrank were placed in the "deteriorated" group. A change in cluster area of less than 25 cells was categorized as "no change."

Using this criterion, 21 patients were identified in the "improved" group, 23 in the "deteriorated" group, and 13 patients exhibited no clear change in voice quality.

TABLE 5.
Statistics Comparing PT and TT

Metric	Mean change for patients with PT	Mean change for patients with TT	Mean difference in the changes	*P*-value	95% CI
Voice range change	− 47.95 ± 87.46	− 58.25 ± 151.58	10.30	0.78	(− 61.43, 82.03)
VHI	2.3 ± 7.86	5.6 ± 7.69	− 3.30	0.23	[− 5.51, − 1.09]
Crest Factor	− 0.05 ± 0.06	− 0.00 ± 0.10	− 0.05	0.24	[− 0.08, − 0.02]
SB (dB)	− 1.68 ± 2.20	0.29 ± 2.48	− 2.00	*	[− 2.80, − 1.20]
CPPs (dB)	− 0.14 ± 0.53	0.09 ± 1.33	− 0.24	0.59	[− 0.58, 0.11]
CSE	0.35 ± 0.72	− 0.04 ± 1.16	0.39	*	[0.10, 1.36]
Q_{Δ}	− 0.40 ± 1.01	− 0.2 ± 1.15	− 0.16	0.64	[− 0.92, 0.51]
Qci	0.02 ± 0.02	0.00 ± 0.04	0.02	*	[0.01, 0.04]

Metric differences were calculated as post minus pre across subjects. CI differences were calculated as PT minus TT group.
*Significant ($P < 0.05$).
No asterisk: not significant.

TABLE 6.
Statistics Comparing "Improved" and "Deteriorated" Groups

Metric	Mean change of improved group	Mean change of deteriorated group	Mean difference in the changes	*P*-value	95% CI
Voice range change	10.52 ± 170.96	− 80.48 ± 132.04	91.0 ± 46.37	0.057	(0.12, 181.88)
VHI	3.33 ± 4.91	5.70 ± 9.13	− 2.36 ± 2.18	0.30	(− 6.6, 1.9)
Crest Factor	0.04 ± 0.12	− 0.06 ± 0.10	0.09 ± 0.03	**	(0.03, 0.16)
SB (dB)	1.01 ± 2.73	− 1.57 ± 2.64	2.64 ± 0.81	**	(1.05, 4.23)
CPPs (dB)	0.73 ± 1.19	− 0.72 ± 0.95	1.5 ± 0.33	***	(0.86, 2.14)
CSE	− 0.71 ± 1.11	0.80 ± 0.61	− 1.5 ± 0.27	***	(− 2.04, − 0.97)
Q_Δ	− 0.23 ± 1.31	− 0.46 ± 0.97	0.23 ± 0.35	0.52	(− 0.46, 0.91)
Qci	− 0.01 ± 0.03	0.01 ± 0.03	− 0.02 ± 0.01	*	(− 0.04, − 0.002)

The first column is the parameters changes (post-pre), with differences calculated as "improved" minus "deteriorated" group across subjects.
***Highly significant ($P < 0.001$).
**Very significant ($P < 0.01$).
*Significant ($P < 0.05$).
No asterisk: not significant.

From Table 6, as expected, we can observe that the "improved" group has a positive voice range change of around 10.52 and the "deteriorated" group shows a negative voice range change of − 80.48.

Both groups have increased VHI postoperation, though the difference between two groups is not significant ($P > 0.05$). This implies that, on average, both groups felt similarly about the impact of their voice issues on their daily life after the operation.

The acoustic metrics, except Q_Δ, show statistically significant differences between two groups. In general, it indicates that the "improved" group is characterized as higher Crest Factor, SB, CPPs and lower CSE and Q_{ci} postoperation.

We then apply a chi-square test to determine if the type of surgery (total vs partial thyroidectomy) influenced the formulation of two groups. Though a higher proportion of patients in the "improved" group underwent a TT compared to the "deteriorated" group (17 out of 21 in the "improved" group vs 15 out of 23 in the "deteriorated" group), the chi-square test shows this difference is not statistically significant (*P*-value = 0.23).

Bayesian Information Criterion

The Bayesian Information Criterion (BIC) is a criterion for model selection among a set of models. It is based on the likelihood function, and it introduces a penalty term for the number of parameters in the model, aiming to avoid overfitting.

Twenty-three out of 57 patients had a minimum BIC value at k = 2. The typical BIC curve for them is shown in Figure 19. It suggests that a two-cluster solution provides the best balance between goodness-of-fit and model complexity for those instances. This BIC is calculated only within one subject for his/her pre- and post-operative data.

However, the BIC curves of the rest are approximately a straight line across different values of k. It implies that none of the tested numbers of cluster provides a significantly better fit than the others and that the data for these patients do not have distinct or separable groupings.

Interestingly, when analyzing the entire dataset, an 8-cluster solution is suggested as optimal. This might be because, when considering all patients together, the data encompasses a broader range of variability, patterns, or subgroups, necessitating a higher number of clusters to capture the inherent structures effectively. This is also the case for subset of the patients in the TT group. However, for the PT group, the optimal clustering solution suggests 3 or 4 clusters (Figure 20). This difference from the entire group and the TT group can be attributed to the fact that the PT group might have more homogeneity or specific patterns that can be better captured with fewer clusters.

This difference in the optimal number of clusters could result from the fact that a PT involves the removal of only part of the thyroid, which might cause less disruption to nearby structures like the RLN, which controls the vocal folds. This could lead to more uniform voice signals.

DISCUSSION

The predominant effect observed postsurgery is a reduction in voice range, particularly in f_o and SPL, already evidenced in the literature.[66] These quantitative changes have practical implications for patients, impacting their vocal capabilities and, by extension, their daily communications and overall quality of life.

Some voice quality metrics remained virtually unchanged postintervention, others exhibited subtle yet statistically significant alterations. For example, the SB and Q_Δ metrics decreased, and the CPPs and CSE metrics increased. The use of K-means clustering of the metrics invokes more distinct patterns of postsurgery changes, with most patients exhibiting characteristics that fit into either a "stable" or "unstable" cluster. The production of voice involves multiple processes, and the changes in voice quality may vary across different types of phonations. An isolated metric represents a narrower aspect of the voice quality, than

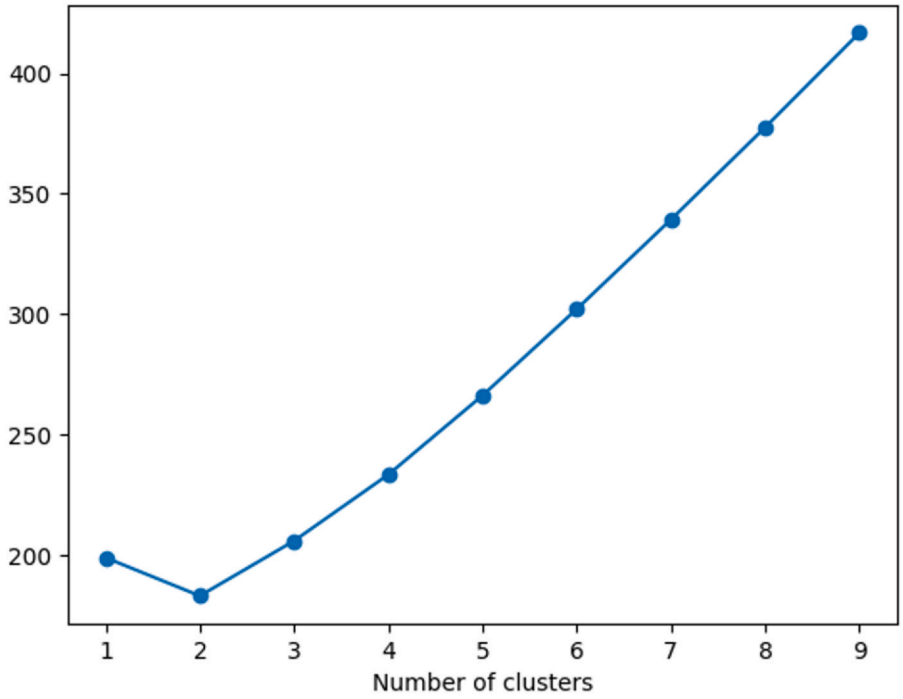

FIGURE 19. BIC values for patient No. 15. BIC, Bayesian Information Criterion.

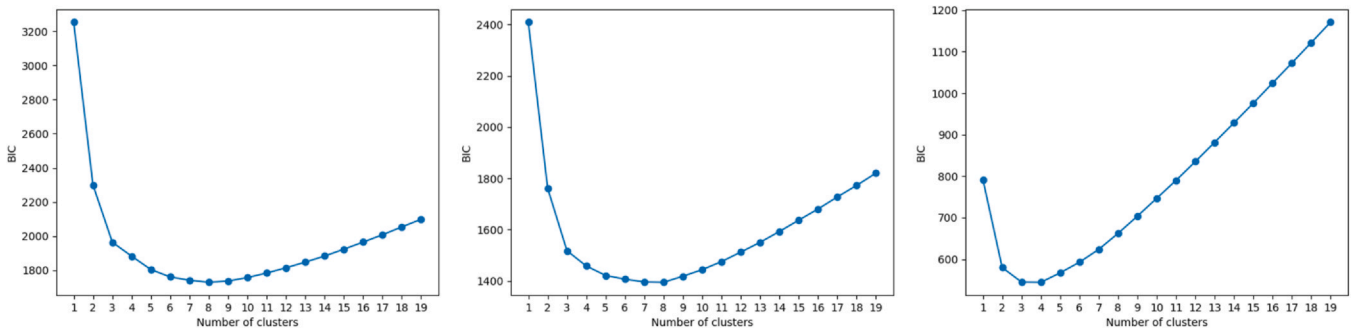

FIGURE 20. BIC curves for left, all patients; mid, all patients with total thyroidectomy; right, all patients with partial thyroidectomy. BIC, Bayesian Information Criterion.

several metrics clustered into phonation types. The present set of metrics is the set of those that were on hand at the time of the study, and is not necessarily an optimal one; but one has to start somewhere. The choice of metrics will ideally depend on the research question. It is possible that there exist effects of TT and PT on voice quality after surgery that are not well represented even by the six metrics employed here.

The classification of vocal productions into a few discrete clusters discards information on effect sizes, since even a small change in just one metric could incur a switch to another cluster. This disadvantage is however offset by using voice maps, with a fairly large number of cells being compared, and with each cell containing an average of many observations. The number of changed cells then gives a relative assessment of the effect size that is conceptually easy to interpret. In the present study, we chose as the outcome the proportion of cells in the overlapping area whose classification changed, from pre- to post-intervention. There are more ambitious ways of judging the effect sizes, such as computing the distances in centroid space between data points pre- and post-intervention[67] but these are much more difficult to explain and to interpret.

VHI correlated primarily with voice range. It did not correlate with the individual metrics, but it was correlated to the binary "stable"/"unstable" clustering of parameters. That is, the more the voice resides in the unstable region, the higher the likelihood of a higher VHI after surgery. This suggests again that while individual metrics might not offer predictive power, a more integrative view that considers clusters of metrics can provide insights into post-operative voice quality. In certain cases, patients may exhibit higher VHI scores after surgery, while paradoxically demonstrating an improvement in voice quality.

This discrepancy may be attributable to individual psychological factors influencing patient self-assessment. In such cases, the voice quality map may serve as an objective benchmark tool to assisting in the reconciliation of patient perceptions with positive surgical outcomes.

Grouping patients by their diagnoses, such as nodules or cancer, reveals limited effects due to small sample sizes within each group. Vocal changes appear to be primarily influenced by treatment interventions rather than the diagnosed conditions themselves.

While we did not conduct a quantitative analysis of trauma due to its rarity among the majority of participants, we attempted to categorize all instances of damage for analysis. After categorizing patients based on reported peri-operative muscle and nerve damage and evaluating their voice metrics, however, we observed no statistically significant differences ($P > 0.05$) between the groups with reported damage (N = 11) and those without (N = 46), both in individual and aggregated metrics.

The laryngological assessments yielded similar results. When comparing these two groups with and without laryngological disorders for intergroup differences, we found no significant differences ($P > 0.05$) across all individual metrics or metric clusters. Still, pathological alterations might not always manifest through voice production alone.

This outcome suggests that surgical results remain consistent irrespective of peri-operative damage or laryngological disorders, possibly attributed to the roughly 2-month recovery interval before postoperative assessment, which may allow sufficient time for vocal rehabilitation or healing. Given the statistics, we cannot make the conclusion that any of the vocal changes resulted from the laryngological or traumatic types.

The type of surgery did influence voice quality. The PT group generally showed "improved" voice metrics and a somewhat larger voice range (10 ST × dB), with fewer complaints and a slightly better-perceived voice quality. However, these differences were not statistically significant. Theoretically and as shown in literature,[68,69] these differences nevertheless align with the variations in surgical procedures, where the PT group involves less organs removal and damage. This lack of significance might be because the two groups are not evenly distributed. In general, the PT group shows greater voice homogeneity and a smaller group size, which is why the BIC indicated a smaller number of clusters.

One limitation of the experiment is its retrospective nature. While this major data set with audio, EGG and much else is an unusual and significant resource, the recordings were somewhat contaminated, in that the EGG signals and sometimes microphone signals often were noisy. As a result, we had to apply a rather high degree of noise reduction to the EGG, as well as other data-cleaning processes, which improved matters somewhat, but still yielded less than ideal measurements. Also, the experimental design was not originally intended for making voice maps; hence the overlapping sections we exploited were more serendipitous than controlled. The patients were not asked to exercise their entire voice range, as for making conventional voice range profile, which would be too challenging for the patients due to anxiety or discomfort. A task protocol could be designed so as to elicit a greater amount of voice range overlap pre- and post-intervention, which would increase the amount of data available for comparisons.[70]

There were significant individual differences in the experiment, with many variations in VHI caused by factors unrelated to voice quality, such as diagnostic results, psychological factors, on-the-spot condition, the response to the surgery and their recovery process. It is difficult to thoroughly understand each individual's condition. Hence, we hope that the conclusions drawn are specific to the recordings themselves, which also implies that a tailored analysis for each patient is required.

The timing of the postoperative assessments can further influence the observed outcomes and thus needs to be controlled.

CONCLUSION

This study looked for possible changes in voice quality (rather than in range) across partial or TT, using no less than six acoustic and EGG metrics, and analyzing *all* phonatory periods across several tasks using the voice-mapping technique. We therefore believe that the main finding, that the intervention effects on individual voice quality metrics at the group level are too small to be clinically relevant, is a very robust finding. At the same time, K-means clustering of these metrics revealed clear effects at the group level, related mostly to an increased degree postintervention of voice perturbations (CPPs and CSE). The instances of per-operative trauma were however too few for us to relate those effects to any particular trauma.

The type of surgery also played a role in voice outcomes. Notably, the PT group, which involves lesser tissue removal, generally exhibited better voice metric outcomes than did the TT group, though not statistically significant. Looking ahead, future studies in this vein should probably concentrate on specific voice problems as they arise after thyroidectomy, and assess these on an individual basis.

Declaration of Competing Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper. Neither financial nor personal relationships have influenced the work reported in this document. This includes but is not limited to, consultancies, employment, advocacies, stock ownership, honoraria, paid expert testimony, patent applications/registrations, and grants or other funding.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the first author used ChatGPT 4.0 in order to improve language and readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgments

Dr. Sébastien Guigard, Dr. Hélène Blaise, Dr. Jocelyne Sarfati and Claire Lalevée-Huart are gratefully acknowledged for performing important tasks in building the data set. Dr. Silvia Capobianco offered insightful comments on and suggestions for the manuscript.

Appendix A

The VHI-10 was given in French. The instructions given were as follows: "You are required to respond to each statement by selecting a single checkbox each time." The available response options were as follows: *J* = Never (*Jamais*); P*J* = Almost Never (*Presque Jamais*); P = Sometimes (*Parfois*); PT = Almost Always (*Presque Toujours*); T = Always (*Toujours*). Each response option corresponds to a specific point value: *J* = 0; P*J* = 1; P = 2; PT = 3; T = 4. The total score ranges from 0 to 40, obtained by summing up the scores of the 10 individual items.

Voice Handicap Index (VHI).

Instructions: These are statements that many people have used to describe their voices and the effects of their voices on their lives. Circle the response that indicates how frequently you have the same experience.

0 = Never 1 = Almost Never 2 = Sometimes 3 = Almost Always 4 = Always

Part I: Functional						
F1	My voice makes it difficult for people to hear me.	0	1	2	3	4
F2	People have difficulty understanding me in a noisy room.	0	1	2	3	4
F3	My family has difficulty hearing me when I call them throughout the house.	0	1	2	3	4
F4	I use the phone less often than I would like to.	0	1	2	3	4
F5	I tend to avoid groups of people because of my voice.	0	1	2	3	4
F6	I speak with friends, neighbors, or relatives less often because of my voice.	0	1	2	3	4
F7	People ask me to repeat myself when speaking face-to-face.	0	1	2	3	4
F8	My voice difficulties restrict personal and social life.	0	1	2	3	4
F9	I feel left out of conversations because of my voice.	0	1	2	3	4
F10	My voice problem causes me to lose income.	0	1	2	3	4
Part II: Physical						
P1	I run out of air when I talk.	0	1	2	3	4
P2	The sound of my voice varies throughout the day.	0	1	2	3	4
P3	People ask, "What's wrong with your voice?"	0	1	2	3	4
P4	My voice sounds creaky and dry.	0	1	2	3	4
P5	I feel as though I have to strain to produce voice.	0	1	2	3	4
P6	The clarity of my voice is unpredictable.	0	1	2	3	4
P7	I try to change my voice to sound different.	0	1	2	3	4
P8	I use a great deal of effort to speak.	0	1	2	3	4
P9	My voice is worse in the evening.	0	1	2	3	4
P10	My voice "gives out" on me in the middle of speaking.	0	1	2	3	4
Part III: Emotional						
E1	I am tense when talking to others because of my voice.	0	1	2	3	4
E2	People seem irritated with my voice.	0	1	2	3	4
E3	I find other people don't understand my voice problem.	0	1	2	3	4
E4	My voice problem upsets me.	0	1	2	3	4
E5	I am less outgoing because of my voice problem.	0	1	2	3	4
E6	My voice makes me feel handicapped.	0	1	2	3	4
E7	I feel annoyed when people ask me to repeat.	0	1	2	3	4
E8	I feel embarrassed when people ask me to repeat.	0	1	2	3	4
E9	My voice makes me feel incompetent.	0	1	2	3	4
E10	I am ashamed of my voice problem.	0	1	2	3	4

.

Appendix B

Number	Task	Description
1	Text Reading	Patients read a provided text excerpt from Antoine de St-Exupéry's book "Le Petit Prince."
2	Spoken Sequence "pɛpɛpɛpɛ"	Patients read a series of syllables ("pɛ") three times at different intensities: normal, soft, and loud, using an intraoral pressure measurement tube.
3	Sustained Sound: "s"	Patients take a breath and produce a sustained /s/ sound for as long as possible, repeating the exercise three times.
4	Sustained Sound: "z"	Patients take a breath and produce a sustained /z/ sound for as long as possible, repeating the exercise three times.
5	Sustained Sound: "e"	Patients take a breath and produce a sustained /e/ sound for as long as possible, repeating the exercise three times.

Number	Task	Description
6	Gentle Rising-Falling Siren	Patients produce gentle rising and falling sirens (glissandi) in a soft (piano) manner.
7	Loud Rising-Falling Siren	Patients produce rising and falling sirens (glissandi) in a loud (forte) manner.
8	"Joyeux Anniversaire" - Free Tonality and Tempo	Patients sing the "Joyeux Anniversaire" song, first slowly and then faster.
9	"Joyeux Anniversaire" - Karaoke	Patients sing the "Joyeux Anniversaire" song with a karaoke version. The tempo starts slow and then becomes faster.
10	"Joyeux Anniversaire" - Karaoke with "pɛpɛpɛ"	Patients sing the "Joyeux Anniversaire" song replacing the lyrics with "pɛpɛpɛ," using the intraoral pressure measurement tube.
11	Sung Sequence "pɛpɛpɛpɛpɛ" - Do3 (C4)	Patients sing the syllables "pɛ pɛ pɛ pɛ pɛ" three times at different intensities: normal, soft, and loud, using the intraoral pressure measurement tube.
12	Sung Sequence "pɛpɛpɛpɛpɛ" - Mi3 (E4)	Patients sing the syllables "pɛ pɛ pɛ pɛ pɛ" three times at different intensities: normal, soft, and loud, using the intraoral pressure measurement tube.
13	Sung Sequence "pɛpɛpɛpɛpɛ" - Sol3 (G4)	Patients sing the syllables "pɛ pɛ pɛ pɛ pɛ" three times at different intensities: normal, soft, and loud, using the intraoral pressure measurement tube.
14	Sung Sequence "pɛpɛpɛpɛpɛ" - Do4 (C5)	Patients sing the syllables "pɛ pɛ pɛ pɛ pɛ" three times at different intensities: normal, soft, and loud, using the intraoral pressure measurement tube.

The original tasks recorded, of which only Task 1, 2, 5, 6, 7, 8, 11-14 are selected for this study.

References

1. Patel KN, Yip L, Lubitz CC, et al. The American Association of Endocrine Surgeons Guidelines for the definitive surgical management of thyroid disease in adults. *Ann Surg.* 2020;271:e23–e57.
2. Fortuny JV, Guigard S, Karenovics W, et al. Surgery of the thyroid: recent developments and perspective. *Swiss Med Wkly.* 2015;145:w14144.
3. Rodriguez A, Hans S, Lechien JR. Post-thyroidectomy voice and swallowing disorders and association with laryngopharyngeal reflux: a scoping review. *Laryngosc Investig Otolaryngol.* 2023;8:140–149.
4. Stojadinovic A, Shaha AR, Orlikoff RF, et al. Prospective functional voice assessment in patients undergoing thyroid surgery. *Ann Surg.* 2002;236:823–832.
5. Shin YJ, Hong KH. Cepstral analysis of voice in patients with thyroidectomy. *Clin Exp Otorhinolaryngol.* 2016;9:157–162.
6. Wojtczak B, Sutkowski K, Kaliszewski K, et al. Voice quality preservation in thyroid surgery with neuromonitoring. *Endocrine.* 2018;61:232–239.
7. Kim GJ, Bang J, Shin HI, et al. Persistent subjective voice symptoms for two years after thyroidectomy. *Am J Otolaryngol.* 2023;44:103820.
8. Haddou N, Idrissi N, Ben Jebara S. Analysis of voice quality after thyroid surgery. *J Voice.* 2023;S0892-1997(23):00208-4. https://doi.org/10.1016/j.jvoice.2023.06.027. Epub ahead of print. PMID: 37612171.
9. Cai H, Ternström S. Mapping phonation types by clustering of multiple metrics. *Appl Sci.* 2022;12:12092.
10. Hwan Hong K, Ye M, Mo Kim Y, et al. The role of strap muscles in phonation—In vivo caninelaryngeal model. *J Voice.* 1997;11:23–32.
11. Kark AE, Kissin MW, Auerbach R, et al. Voice changes after thyroidectomy: role of the external laryngeal nerve. 1984;289:4.
12. Pereira JA, Girvent M, Sancho JJ, et al. Prevalence of long-term upper aerodigestive symptoms after uncomplicated bilateral thyroidectomy. *Surgery.* 2003;133:318–322.
13. Shimokojin T, Takenoshita M, Sakai T, et al. Vocal cordal bowing as a cause of long-lasting hoarseness after a few hours of tracheal intubation. *Anesthesiology.* 1998;89:785–787.
14. Çalışkan M, Demirci T, Cengiz H. Evaluation of voice quality in primary hyperparathyroidism patients undergoing minimally invasive parathyroid surgery. *Cir Cir.* 2022;90:45–51.
15. Henry LR, Helou LB, Solomon NP, et al. Functional voice outcomes after thyroidectomy: an assessment of the Dsyphonia Severity Index (DSI) after thyroidectomy. *Surgery.* 2010;147:861–870.
16. Veldova Z, Holy R, Rotnagl J, et al. Influence of recurrent laryngeal nerve transient unilateral palsy on objective voice parameters and on voice handicap index after total thyroidectomy (including thyroid carcinoma). *Int J Environ Res Public Health.* 2021;18:4300.
17. Scerrino G, Inviati A, Di Giovanni S, et al. Esophageal motility changes after thyroidectomy; possible associations with postoperative voice and swallowing disorders: preliminary results. *Otolaryngol Head Neck Surg.* 2013;148:926–932.
18. Lombardi CP, Raffaelli M, D'Alatri L, et al. Voice and swallowing changes after thyroidectomy in patients without inferior laryngeal nerve injuries. *Surgery.* 2006;140:1026–1034.
19. Randolph GW, Dralle H, Abdullah H, et al. Electrophysiologic recurrent laryngeal nerve monitoring during thyroid and parathyroid surgery: international standards guideline statement. *Laryngoscope.* 2011;121(suppl 1):S1–16.
20. Terris DJ, Chaung K, Duke WS. Continuous vagal nerve monitoring is dangerous and should not routinely be done during thyroid surgery. *World J Surg.* 2015;39:2471–2476.
21. Suh I, Yingling C, Randolph GW, et al. A novel method of neuromonitoring in thyroidectomy and parathyroidectomy using transcutaneous intraoperative vagal stimulation. *JAMA Surg.* 2016;151:290–292.
22. Cirocchi R, Arezzo A, D'Andrea V, et al. Intraoperative neuromonitoring versus visual nerve identification for prevention of recurrent laryngeal nerve injury in adults undergoing thyroid surgery. *Cochrane Database Syst Rev.* 2019;1:Cd012483.
23. Davey MG, Cleere EF, Lowery AJ, et al. Intraoperative recurrent laryngeal nerve monitoring versus visualisation alone - a systematic review and meta-analysis of randomized controlled trials. *Am J Surg.* 2022;224:836–841.
24. Dralle H, Sekulla C, Haerting J, et al. Risk factors of paralysis and functional outcome after recurrent laryngeal nerve monitoring in thyroid surgery. *Surgery.* 2004;136:1310–1322.
25. Chan WF, Lang BH, Lo CY. The role of intraoperative neuromonitoring of recurrent laryngeal nerve during thyroidectomy: a comparative study on 1000 nerves at risk. *Surgery.* 2006;140:866–872.
26. Al-Qurayshi Z, Randolph GW, Alshehri M, et al. Analysis of variations in the use of intraoperative nerve monitoring in thyroid surgery. *JAMA Otolaryngol Head Neck Surg.* 2016;142:584–589.
27. Barczyński M, Konturek A, Cichoń S. Randomized clinical trial of visualization versus neuromonitoring of recurrent laryngeal nerves during thyroidectomy. *Br J Surg.* 2009;96:240–246.
28. Dionigi G, Boni L, Rovera F, et al. Neuromonitoring and video-assisted thyroidectomy: a prospective, randomized case-control evaluation. *Surg Endosc.* 2009;23:996–1003.
29. Pisanu A, Porceddu G, Podda M, et al. Systematic review with meta-analysis of studies comparing intraoperative neuromonitoring of

recurrent laryngeal nerves versus visualization alone during thyroidectomy. *J Surg Res.* 2014;188:152–161.
30. Sarı S, Erbil Y, Sümer A, et al. Evaluation of recurrent laryngeal nerve monitoring in thyroid surgery. *Int J Surg.* 2010;8:474–478.
31. Vasileiadis I, Karatzas T, Charitoudis G, et al. Association of intraoperative neuromonitoring with reduced recurrent laryngeal nerve injury in patients undergoing total thyroidectomy. *JAMA Otolaryngol Head Neck Surg.* 2016;142:994–1001.
32. Bergenfelz A, Salem AF, Jacobsson H, et al. Risk of recurrent laryngeal nerve palsy in patients undergoing thyroidectomy with and without intraoperative nerve monitoring. *Br J Surg.* 2016;103:1828–1838.
33. Brajcich BC, McHenry CR. The utility of intraoperative nerve monitoring during thyroid surgery. *J Surg Res.* 2016;204:29–33.
34. Thomusch O, Sekulla C, Walls G, et al. Intraoperative neuromonitoring of surgery for benign goiter. *Am J Surg.* 2002;183:673–678.
35. Sinagra DL, Montesinos MR, Tacchi VA, et al. Voice changes after thyroidectomy without recurrent laryngeal nerve injury. *J Am Coll Surg.* 2004;199:556–560.
36. Lombardi CP, Raffaelli M, De Crea C, et al. Long-term outcome of functional post-thyroidectomy voice and swallowing symptoms. *Surgery.* 2009;146:1174–1181.
37. de Pedro Netto I, Fae A, Vartanian JG, et al. Voice and vocal self-assessment after thyroidectomy. *Head Neck.* 2006;28:1106–1114.
38. Debruyne F, Ostyn F, Delaere P, et al. Acoustic analysis of the speaking voice after thyroidectomy. *J Voice.* 1997;11:479–482.
39. Van Lierde K, D'Haeseleer E, Wuyts FL, et al. Impact of thyroidectomy without laryngeal nerve injury on vocal quality characteristics: an objective multiparameter approach: Impact of Thyroidectomy on Vocal Quality. *Laryngoscope.* 2010;120:338–345.
40. Noel JE, Kligerman MP, Megwalu UC. Intraoperative corticosteroids for voice outcomes among patients undergoing thyroidectomy: a systematic review and meta-analysis. *Otolaryngol Head Neck Surg.* 2018;159:811–816.
41. Mcivor NP, Flint DJ, Gillibrand J, et al. Thyroid surgery and voice-related outcomes. *Aust N Z J Surg.* 2000;70:179–183.
42. Hong KH, Kim YK. Phonatory characteristics of patients undergoing thyroidectomy without laryngeal nerve injury. *Otolaryngol Head Neck Surg.* 1997;117:399–404.
43. Vicente DA, Solomon NP, Avital I, et al. Voice outcomes after total thyroidectomy, partial thyroidectomy, or non-neck surgery using a prospective multifactorial assessment. *J Am Coll Surg.* 2014;219:152–163.
44. Starmer HM, Tippett DC, Webster KT. Effects of laryngeal cancer on voice and swallowing. *Otolaryngol Clin N Am.* 2008;41:793–818.
45. Shah RK, Engel SH, Choi SS. Relationship between voice quality and vocal nodule size. *Otolaryngol Head Neck Surg.* 2008;139:723–726.
46. Mohammadzadeh A, Heydari E, Azizi F. Speech impairment in primary hypothyroidism. *J Endocrinol Investig.* 2011;34:431–433.
47. Pabon P, Ternström S. Feature maps of the acoustic spectrum of the voice. *J Voice.* 2020;34.161.e161–161.e126.
48. Brockmann-Bauser M, Bohlender JE, Mehta DD. Acoustic perturbation measures improve with increasing vocal intensity in individuals with and without voice disorders. *J Voice.* 2018;32:162–168.
49. Naomi Anna Iob LH, Ternström S, Cai H, et al. Effects of speech characteristics on electroglottographic and instrumental acoustic voice analysis metrics in women with structural dysphonia before and after treatment. *J Speech Lang Hear Res.* 2024.
50. Angelos P. Recurrent laryngeal nerve monitoring: state of the art, ethical and legal issues. *Surg Clin N Am.* 2009;89:1157–1169.
51. Grubbs EG, Rich TA, Li G, et al. Recent advances in thyroid cancer. *Curr Probl Surg.* 2008;45(3):156–250.
52. Rosen CA, Lee AS, Osborne J, et al. Development and validation of the voice handicap index-10. *Laryngoscope.* 2004;114:1549–1556.
53. Barsties B, Kropp J, Dicks P, et al. [Reliability and Validity of the "Voice Handicap Index (VHI) adapted to the singing voice"]. *Laryngorhinootologie.* 2015;94:441–446.
54. Forti S, Amico M, Zambarbieri A, et al. Validation of the Italian Voice Handicap Index-10. *J Voice.* 2014;28:263.e217–263.e222.
55. Ghio A., Teston B. Evaluation of the acoustic and aerodynamic constraints of a pneumotachograph for speech and voice studies. Proceedings of International Conference on Voice Physiology and Biomechanics. 2004;55–58.
56. Ghio A., Teston B. Caractéristiques de la dynamique d'un Pneumotachographe pour l'étude de la production de la parole: aspects acoustiques et aérodynamique. 2002:24–27.
57. Ternström S. Update 3.1 to FonaDyn: a system for real-time analysis of the electroglottogram, over the voice range. *SoftwareX.* 2024;26:74–80.
58. Ternström S. FonaDyn 3.1.0 Handbook; 2023:1-96.
59. Ternstrom S. Normalized time-domain parameters for electroglottographic waveforms. *J Acoust Soc Am.* 2019;146:EL65.
60. Richman JS, Moorman JR. Physiological time-series analysis using approximate entropy and sample entropy. *Am J Physiol Heart Circ Physiol.* 2000;278:H2039–2049.
61. Selamtzis A, Ternstrom S. Analysis of vibratory states in phonation using spectral features of the electroglottographic signal. *J Acoust Soc Am.* 2014;136:2773–2783.
62. Fant G. The LF-model revisited. Transformations and frequency domain analysis. *STL-QPSR.* 1995;36:119–156.
63. Heman-Ackah YD, Sataloff RT, Laureyns G, et al. Quantifying the cepstral peak prominence, a measure of dysphonia. *J Voice.* 2014;28:783–788.
64. Awan SN, Solomon NP, Helou LB, et al. Spectral-cepstral estimation of dysphonia severity: external validation. *Ann Otol Rhinol Laryngol.* 2013;122:40–48.
65. Arthur D, Vassilvitskii S. *k-means++: the advantages of careful seeding. Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. New Orleans, LA: Society for Industrial and Applied Mathematics.; 2007:1027–1035.
66. Lang BH, Wong CK, Ma EP. A systematic review and meta-analysis on acoustic voice parameters after uncomplicated thyroidectomy. *Laryngoscope.* 2016;126:528–537.
67. Ternström S, D'Amario S, Selamtzis A. Effects of the lung volume on the electroglottographic waveform in trained female singers. *J Voice.* 2020;34.485.e481–485.e421.
68. Ryu J, Ryu YM, Jung YS, et al. Extent of thyroidectomy affects vocal and throat functions: a prospective observational study of lobectomy versus total thyroidectomy. *Surgery.* 2013;154:611–620.
69. Nam IC, Bae JS, Lee SH, et al. Prospective voice assessment after uncomplicated thyroidectomy: a comprehensive analysis of a single centre experience. *Clin Otolaryngol.* 2023;48:39–49.
70. Kim JW, Kwon SB. Development of a simple measurement method for voice range profile examination 1. *J Speech.* 2023;32:001–008.