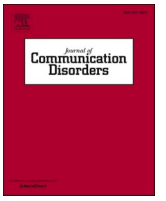
Journal of Communication Disorders 101 (2023) 106292

Contents lists available at ScienceDirect

Journal of Communication Disorders

journal homepage: www.elsevier.com/locate/jcomdis



Objective speech outcomes after surgical treatment for oral cancer: An acoustic analysis of a spontaneous speech corpus containing 32.850 tokens

Thomas B. Tienkamp [a,*], Rob J.J.H. van Son [b,c], Bence Mark Halpern [b,c,d]

[a] Center for Language and Cognition Groningen, University of Groningen, Oude Kijk in 't Jatstraat 26, 9712 EK, Groningen, The Netherlands
[b] Amsterdam Center for Language and Communication, University of Amsterdam, Spuistraat 134, 1012 VB Amsterdam, The Netherlands
[c] Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX, Amsterdam, The Netherlands
[d] Multimedia Computing Group, Delft University of Technology, Mekelweg 4, 2628 CD, Delft, The Netherlands

ARTICLE INFO

Keywords:
Acoustic analysis
Oral cancer
Tongue reconstruction
Spontaneous speech

ABSTRACT

Introduction: Surgical treatment for oral cancer leads to lasting changes of the vocal tract and individuals treated for oral cancer (ITOC) often experience speech problems. The purpose of this study was to analyse the acoustic properties of the spontaneous speech of individuals who were surgically treated for oral cancer. It was investigated (1) how key spectral measures of articulation change post-treatment; (2) whether changes are more related to target manner or place of articulation; and (3) how spectral measures develop at various time points following treatment.
Method: A corpus consisting of 32.850 tokens was constructed by manually segmenting the speech of five (four female - one male) American English speaking ITOC. General acoustic characteristics (duration and spectral tilt), plosives (burst frequency), fricatives (centre of gravity and spectral skewness), and vowels (F1 and F2) were analysed using linear mixed effects regression and compared to control speech. Moreover, a within speaker analysis was performed for speakers with multiple recordings.
Results: Manner of articulation is more predictive of post-treatment changes than place of articulation. Compared to controls, ITOC produced the fricatives /f, v, θ, ð, s, z, ʃ, ʒ/ with a lower centre of gravity while no differences were found for plosives and vowels. Longitudinal analyses show high within-speaker variation, but general improvements one-year post-treatment.
Conclusions: Surgical oral cancer treatment changes the spectral properties of speech. Fricatives with varying manner of articulations were distorted, suggesting that manner of articulation is more predictive than place of articulation in identifying general problem areas for ITOC.

1. Introduction

Tumours in the oral cavity damage various anatomical structures by tumour extension and treatment (De Bruijn et al., 2009). Well known problems for individuals treated for oral cancer (ITOC) include dysphagia, trismus, and challenges in the articulatory domain (Bressmann, 2013). The articulatory problems this group experiences may arise due to various reasons: the tumour's place in the oral cavity, surgical intervention (e.g., hemiglossectomy), or chemoradiation therapy (CRT). In the case of a hemiglossectomy, in which

* Corresponding author at: Oude Kijk in 't Jatstraat 26, 9712 EK Groningen, The Netherlands.
E-mail address: t.b.tienkamp@rug.nl (T.B. Tienkamp).

https://doi.org/10.1016/j.jcomdis.2022.106292
Received 1 April 2022; Received in revised form 21 November 2022; Accepted 27 November 2022
Available online 30 November 2022

half of the tongue is resected, articulatory abilities will inevitably be distorted. This is because the tongue is one of the most important articulators we use in speech production - it acts as an airstream valve by creating narrow or full constrictions through contact with other articulators (Laaksonen et al., 2011). If parts of the tongue are resected, then forming these constrictions may become more difficult due to reduced mobility or missing tissue through which more air can escape, which may impact the intelligibility of the speech output.

Much research has targeted the intelligibility of speech following oral cancer treatment in terms of perceptual evaluations (e.g., Bressmann et al., 2004, 2010). In this case, intelligibility is most often measured using a Likert scale which is typically filled out by speech language pathologists or naive listeners with no clinical or phonetic training. This perception research has contributed greatly to the understanding the impact of tumour site and size (Bressmann et al., 2010; Nicoletti et al., 2004), reconstruction technique (e.g. Bressman et al., 2004; 2010) and tongue mobility (Bressman et al., 2004) on intelligibility. However, perception studies are known to be highly subjective. What sounds intelligible to one rater might sound distorted to another (Ghio et al., 2013; Oates, 2009).

More objective and nuanced accounts of post-treatment changes in speech production may be provided by studying the acoustic properties of speech directly. This method is able to provide a detailed account of specific acoustic parameters whereas much detail may be lost when speech samples are rated on three or four point Likert scales. For example, Takatsu et al. (2017) studied vowel formants in individuals who underwent partial glossectomy and noted that they produced a significantly decreased second formant (F2) for /i/ 5 to 14 days post-treatment. Furthermore, general centralisation of both the first formant (F1) and F2 was also reported for corner vowels, which signals the inability to move the tongue to the required height. This reduced Vowel Space Area (VSA) could reduce intelligibility as the acoustic distance between vowels is reduced, making the vowels less perceptually discriminative. Similar findings were reported by Laaksonen et al. (2010) as individuals who underwent hemiglossectomy produced vowels with less extreme values 1, 6, and 12 months post-treatment. The authors did not find evidence for a reduced VSA, but significant formant changes were observed which suggests that compensatory strategies were used to keep maximal distinctions between vowels. Even though significant improvements are noted post-treatment, acoustic changes in vowel production are long term and remain one year post-treatment. Vowel formants did not return to pre-treatment levels and are still more centralised than they were preoperatively as evidenced by a more posterior tongue position in the case of /a/ (reduced F2) and a lower tongue position in the case of /i/ (increase in F1) (Jacobi et al., 2013; Laaksonen et al., 2010; Takatsu et al., 2017).

Acoustic analyses have also been employed to study post-treatment changes in consonant production. For example, De Bruijn et al. (2009) and Jacobi et al. (2013) studied plosive production (/p, t, tʃ, k/) and found that the burst frequency is reduced following treatment. Moreover, individuals treated for larger tumours had a shorter release of /k/ compared to individuals with smaller tumours (De Bruijn et al., 2009). These lower burst energy levels and shorter release times signal difficulty in building up the necessary pressure to release these plosives. Like with vowels, these changes were long term as burst energy levels did not return to preoperative levels one year post-treatment.

Another class of sounds that pose great difficulty for ITOC, are sibilants (s, z, ʃ, ʒ). Acher et al. (2014), Laaksonen et al. (2011) and Zhou et al. (2011) all observed that the centre of gravity (CoG) of all sibilants significantly decreased following treatment. There was no significant difference between individuals who did and did not receive radiation therapy. Moreover, the spectral energy was localised on the lower frequency bands, and together with a reduced CoG, this points towards a more backed pronunciation (Acher et al., 2014). These acoustic changes were long term for male individuals as they did not return to preoperative levels (Laaksonen et al., 2011). The reduction of the acoustic distinction between /s/, /z/, and /t/ /ʃ/ was short term which implies that, although there were still long term changes, individuals made use of compensatory strategies to clearly maximise the acoustic distance.

Even though acoustic analyses are of great importance as they are able to provide a detailed and objective account of the changes post-treatment, they are scarce. Balaguer et al. (2020) note that at present, 22 acoustic studies have been conducted with ITOC across different languages. Furthermore, a large part of these studies suffer from limitations. For instance, most of these studies used isolated words or sentences for their analysis and often use only a few tokens per type, sometimes only two or three tokens (e.g., De Bruijn et al., 2009; Takatsu et al., 2017). This raises questions about ecological validity for two reasons. First, speech is highly variable and more tokens are needed in order to filter out this variability. Second, single words or sentences might reflect the individual's best effort rather than how they articulate in their everyday communication since ITOC have more time to plan their speech. An additional limitation of the current literature, is that most studies focus on a single group of sounds (e.g., corner vowels or sibilants) which makes it hard to establish whether problems with one class of sounds are associated with problems with another class. Jacobi et al. (2013) provided such an analysis by looking at corner vowels, plosives, fricatives and laterals, but their study focused only on individuals who received CRT as opposed to surgical treatment.

2. The present study

It becomes apparent that a thorough acoustic analysis of multiple sound classes that focuses on spontaneous speech of individuals who were surgically treated for oral cancer is missing. This information could provide a more naturalistic account of how surgical treatment for oral cancer affects everyday speech rather than tightly controlled lab speech. Moreover, a better understanding of acoustic changes could inform speech rehabilitation strategies as current standardised therapies are absent (Blyth et al., 2015). The present study was conducted in order to fill this gap. The research questions can thus be stated as follows:

(1) How do key spectral measures of articulation (e.g., centre of gravity, spectral tilt, and vowel formant frequencies) change in the speech of ITOC post-treatment?
(2) Are changes related to manner and place of articulation (MoA & PoA)?

(3) For the two participants with data at multiple time points post-treatment, how do the above mentioned spectral measures develop at various time points following treatment?

To this end, a speech corpus that contains one hour of running speech of five speakers (approximately 32.850 tokens) was constructed. Acoustic parameters were analysed based on the MoA and PoA while controlling for time after treatment (see Section 3.4). The speech of ITOC was compared to control speech in order to evaluate the speaker's articulatory precision. Table 1 shows the target MoA and PoA of the phonemes under investigation. Laterals and nasals were not included in the analysis. Laterals /r, l/ were excluded as they only have one PoA while the present study investigates interactions between MoA and PoA. Nasals /m, n, ŋ/ were excluded since none of the speakers were treated for a tumour in the sinonasal area or nasopharynx, meaning that we did not expect difficulty in this area whereas problems with plosives and fricatives have been reported in individuals who were treated for lingual tumours. It should be noted that ITOC might not be able to produce these typical targets. Instead, recent articulatory work by Hagedorn et al. (2022) revealed that alveolar targets varied systematically as a function of the target MoA. Therefore, it cannot be ruled out that the PoA of the clinical group manifest themselves in the data in a similar fashion as that of the control speakers. Besides these consonants, the corner vowels /i/, /a/, and /u/ were analysed, too.

Even though the research questions are quite exploratory in their nature, some predictions can still be made. For research question (1), it was hypothesised that oral cancer treatment influences the acoustic properties of the patient's speech. More specifically, it was predicted that there would be more vowel centralisation in ITOC due to a decrease in tongue mobility based on Jacobi et al. (2013) and Takatsu et al. (2017). For plosives, it was predicted that ITOC would produce them with a lower burst frequency (BF) than control speech based on the results of Jacobi et al. (2013) and de Bruin et al. (2009). For sibilants, a lower CoG was expected as well as that the spectral energy would be localised on the lower frequencies (higher skewness) based on the results of Laaksonen et al. (2011), Zhou et al. (2011), and Acher et al. (2014).

For research question (2), based on the correlation between tongue mobility and perceptual quality (Bressmann et al., 2004), it was hypothesised that sounds using the tongue-tip ([+] coronal sounds like alveolars) would be more affected than sounds that do not use the tongue-tip ([-] coronal sounds like bilabials). Therefore, it was predicted that there would be a bigger difference between clinical and control speech in the burst frequencies of alveolar plosives (/t/, /d/) as compared to bilabial plosives (/b/, /p/). The velar plosives (/k/, /g/) may be impaired too as a result of scar tissue in the back of the mouth that hampers velopharyngeal functioning (cf. de Bruin et al., 2009 and Jacobi et al., 2013). Furthermore, it was predicted that there would be a bigger difference in CoG and skewness between dental, alveolar, and post-alveolar fricatives as compared to labio-dental fricatives as the latter are [-] coronal. It was predicted that affricates would pose difficulty for the oral cancer speakers both on the grounds of the manner and place of articulation. No specific predictions were made with regards to voicing. ITOC's main difficulty is thought to be on the articulatory side due to anatomical changes following surgery. The larynx, or the voice box, is mostly spared which differs from individuals with laryngeal cancers (e.g., van Sluis et al., 2019). Still, CRT is known to affect laryngeal functioning and some of the speakers included in the present study have received CRT (cf. Jacobi et al., 2013; Lazarus, 2009). However, null-effects regarding CRT have been documented, too (e.g. Laaksonen et al., 2011; Zhou et al., 2011).

Lastly, for research question (3), it was predicted that for Male01 the burst frequency of plosives would increase post-treatment based on Jacobi et al. (2013) and de Bruin et al. (2009). The vowel centralisation was predicted to improve (i.e., less centralisation), as the VSA increased after a mean rehabilitation period of 50 days (Takatsu et al., 2017). Fricatives were also predicted to improve in the sense that the CoG would increase, and the spectral skewness would decrease based on Laaksonen et al. (2011) and Acher et al. (2014). No clear predictions could be made for the development of Female01 due to the timing of the recordings.

3. Methodology

3.1. Materials and design

Speech recordings were selected from the online platform YouTube and have been presented in full in Halpern et al. (2020).[1] To fit the aims of the present study, a sub-selection was made. ITOC were included if: (1) they were surgically treated for oral cancer, (2) were native speakers of American English, (3) were intelligible enough in order to make phonemic transcriptions, and (4) were of good acoustic quality. A flowchart describing the exclusion criteria can be found in Fig. 1. The corpus used for the current analysis consists of five speakers (four female - one male) and is one hour long.[2] Details of the speakers, such as the duration of recordings per speaker, as well as the number of tokens per recording can be found in Table 2. We use the term 'hemiglossectomy' when half of the tongue is resected, and 'partial glossectomy' when part, but not half, of the lateral tongue is resected. Speakers did not explicitly mention their age, making it impossible to give precise information. To provide a rough estimate, we follow Lachman's (2001) distinction between early (between 20 and 39 years old) and middle adulthood (between 40 and 59 years old). The female speakers were all in early adulthood whereas the male speaker was in middle adulthood. The distribution of each phoneme per speaker is provided in the Supplementary Materials (S-Fig 1**)**. Perceptual quality of the articulation was assessed by an American SLP. The possible ratings included severe (1) to normal (5.0). Ratings were based on the ease at which one could transcribe the utterance, how clear the

[1] The entire corpus can be accessed at: https://doi.org/10.5281/zenodo.3732322

[2] The female-male (sex) distribution was inferred based on the videos. However, it was impossible to infer the gender identity of the clinical speakers as they did not explicitly address this. We were thus unable to verify the gender of every speaker.

Table 1
Place and manner of articulation of analysed phonemes.

Place → Manner↓	Bilabial	Labio- dental	Dental	Alveolar	Post- Alveolar	Velar
Plosive	/b/ /p/			/t/ /d/		/k/ /g/
Fricative		/f/ /v/	/θ/ /ð/	/s/ /z/	/ʃ/ /ʒ/	
Affricate				/tʃ/ /dʒ/		

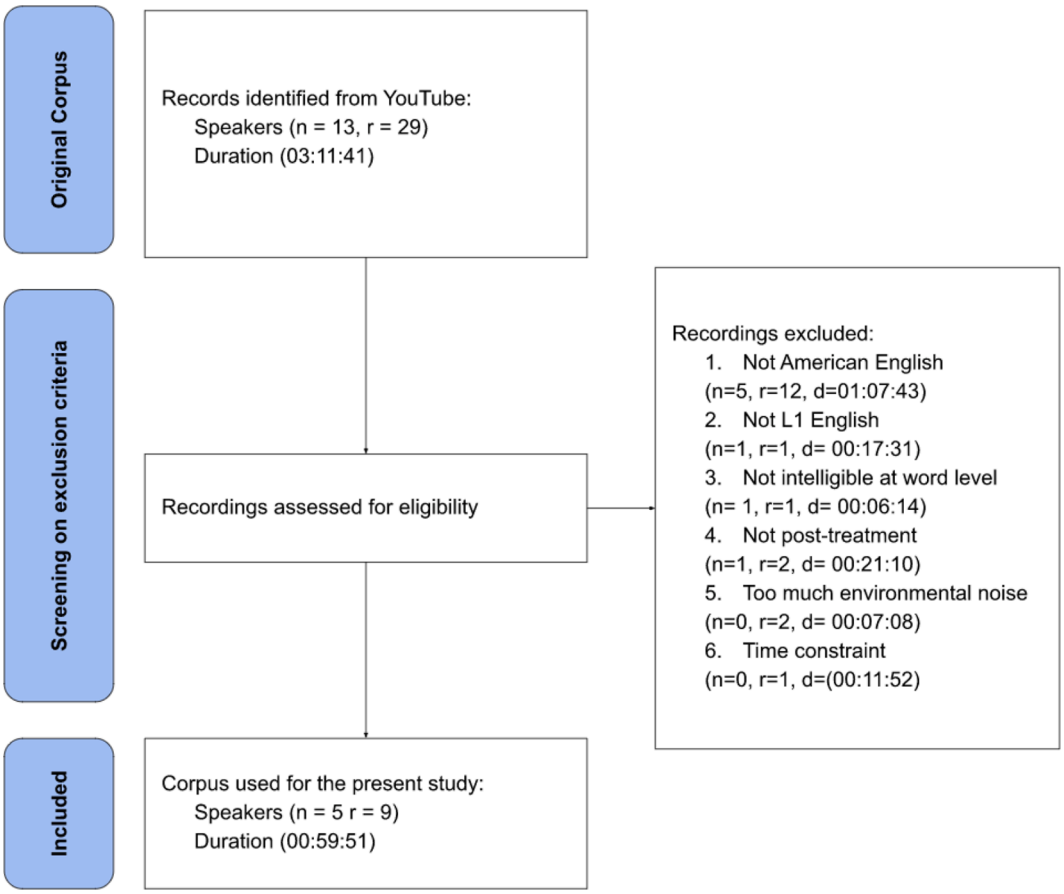


Fig. 1. Flowchart of recordings and people included in the corpus. n = speakers, r = recordings, d = duration which is given in hours:minutes:seconds.

utterance was, and whether substantial changes in terms of speech rate were noticeable. A full description of the instructions can be found in Appendix A. Additionally, the speech rate was automatically measured using the script by De Jong and Wempe (2009) and is specified as the number of syllables per second (clinical speakers: $M = 2.95$, SD = 0.49; control speakers: $M = 3.51$, SD = 0.61).

As no pre-treatment recordings were available for any of the speakers, control speech was selected instead. Control speech was taken from the TIMIT acoustic-phonetic continuous speech corpus Garofolo et al. (1993). The TIMIT corpus is a large corpus that consists of recordings from 630 speakers from the eight major dialect regions of the United States. Ideally, speakers would be matched on year of birth and dialect to ensure that they speak the same diachronic variety of English. However, this was only possible for the male speaker (Male01). Dialect regions were selected on the basis of geographical data provided in the videos. If these data were not available, we selected the region based on perceptual assessment by one of the authors (TT). For the female speakers in our corpus, the youngest speakers from TIMIT were selected to match them as closely as possible. In total, ten speakers per dialect region were selected from TIMIT to match the amount of tokens per clinical speaker and control dialect region, approximately 4.000 tokens per dialect region (11.700 in total). All sentences per speaker were included to maximise the number of tokens for each phoneme. The number of tokens per TIMIT dialect region, and a full list of included TIMIT speakers, can be found in the Supplementary Materials (S-Tables 1-2). It is noted, however, that these baseline recordings need not reflect the pre-operative speech of the individual. Instead, the control speakers provide a baseline of acceptable ranges for the acoustic parameters under scrutiny in the present study.

Table 2
Speaker information. RFFF = radial forearm free flap. CRT = chemoradiation therapy. mm:ss = minutes:seconds. Speech rate = number of syllables/second. Middle = middle adulthood (40–59 years of age). Early = early adulthood (20–39 years of age).

Speaker	Age	Procedure	Place	Reconstruction	CRT	Time-stamps	Duration (mm:ss)	Tokens	Perceptual rating	Speech rate
Male01	Middle	Hemi-glossectomy	Left half removed	RFFF	Yes	2 months	13:55	7.677	5.0	2.81
						4 months	1:25	789	5.0	2.56
						8 months	4:03	2.395	5.0	3.01
						26 months	6:30	3.772	–[a]	3.37
						40 months	3:10	2.352	5.0	3.71
Female01	Early	Partial-glossectomy	Left middle lateral side	–	–	5 days	1:40	691	3.5	2.32
						11 days	3:20	1.536	4.5	2.52
Female02	Early	Hemi-glossectomy	Side not specified[b]	–	Yes	2 months	6:30	3.479	–[2]	3.18
Female03	Early	Partial-glossectomy	Left middle lateral side	–	Yes	12 months	12:51	6.168	4.0	2.63
Female04	Early	Hemi-mandibulectomy	Right mandible	Titanium plate + bone taken from the scapula	–	16–17 months	6:07	3.987	5.0	3.86

[a]As this is a shared recording, no rating was provided.
[b]We refrain from using the video to make an inference regarding the side of the resection as the video might be mirrored by the recording device.

3.2. Phonemic transcription

After the selection of recordings was made, time aligned transcriptions were made using the software ELAN Version 5.4 (ELAN 2020). As the ITOC were intelligible on the word-level, this did not result in difficulty. Then, starting from the baseline transcriptions, time aligned phonemic transcriptions (using SAMPA transcription) were made using WebMAUS (Kisler et al., 2017). Any errors in segmentation or labelling were manually corrected.

3.3. Acoustic analyses

Acoustic analyses were performed on all labelled segments in order to evaluate the speaker's articulatory precision. The following acoustic parameters were measured for consonants: the duration in milliseconds (ms) CoG as measured by the mean frequency of the phone segment in Hertz (Hz) as weighted by the power spectrum with a window of 5 ms; the burst frequency measured as the CoG at the point of highest intensity in Hz; and the spectral skewness. The skewness parameter informs about the right-left asymmetry: the higher the skewness the more the spectral energy is localised on low frequencies (left skewed) (Acher et al., 2014). Lastly, the spectral tilt was measured at the point of highest intensity in a spectral window of 550–5500 Hz and a temporal window of 5 ms, resulting in a logarithmic measure of the dB per decade. Spectral tilt is a measurement of intensity and describes the spread of intensity amongst different frequency bands.

For the corner vowels, we measured the first two formants (F1 and F2) using a Linear Predictive Coding algorithm with a window length of 25 ms and a formant ceiling of 5500 Hz. The vowel was measured at the temporal midpoint to try to account for coarticulation. For the plosives, the centre of the burst frequency was measured. Following Acher et al. (2014), the CoG and the spectral skewness were measured for fricatives. A combination of these measurements were used for the affricates. For all consonants under investigation, the spectral tilt and duration were measured. The programme Praat (Boersma and Weenink, 2020) was used to extract the acoustic measurements. Impossible F1 values for /i/ and /u/ (over 1000 Hz) were excluded from the analysis. This excluded 72 tokens (2.5%) in the group analysis; zero tokens for Female01; and 52 tokens (4.7%) for Male01.

3.4. Statistical analyses

The resulting acoustic measurements were analysed in R version 4.1.2 with linear mixed effects models using the *lmer* function of the *lme4* package 1.1.27.1 (Bates et al., 2015; R Core Team, 2020). *P*-values were approximated with the *lmerTest* package 3.1.3 (Kuznetsova et al., 2017).

To answer research questions (1) and (2), how acoustic parameters change post-treatment and whether these changes are related to either PoA or MoA, we fitted linear mixed effects models with the acoustic parameter as a function of Group (Typical - Clinical), Voicing (Voiced - Voiceless) and the MoA plus PoA combination (e.g., alveolar plosives, alveolar fricatives). The hypothesis model included an interaction between Group and MoA-PoA combination. Using the *anova* function and Akaike Information Criterion (AIC) we compared whether the additional interaction with Voicing improved the model. We took a decrease of 2.0 in AIC as evidence that the inclusion of the interaction improved the model. All numerical predictors were mean centred. All categorical predictors were coded with sum-to-zero orthogonal contrasts: Group (Typical as −0.5, Clinical as +0.5) and Voicing (Voiceless as -0.5, Voiced as +0.5) (Baguley, 2012). To avoid problems of non-convergence as much as possible, the number of maximum iterations was increased to 100, 000 (Powell, 2009). Random intercepts were fitted for individual speakers and time post-treatment because the recordings of the clinical speakers differed with regards to post-treatment time. Again, we used the AIC to determine whether adding a random slope for Voicing for both random intercepts improved the model. We followed the model criticism procedure by Baayen (2008) and assessed our models on autocorrelation, multicollinearity, normality of the residuals, and heteroscedasticity using the *car* package version 3.0.12 (Fox & Weisberg, Fox and Weisberg, 2019). If the model violated one of these checks, we refitted the model with trimmed residuals, preserving at least 95% of our data, or transformed our dependant variable using either a square-root or log transformation. In these instances, the transformation method is mentioned in the text. For the fixed effects, post-hoc comparisons between groups or time-points were performed using the *emmeans* function of version 1.7.2 of the *emmeans* package (Lenth, 2022). However, this was too computationally expensive for the group comparisons due to the number of observations. We therefore refitted these models manually by changing the reference level of the MoA-PoA factorial term in order to get the pairwise comparisons.

To answer research question (3), how spectral measures change over time, a within speaker analysis was carried out for the two speakers of which there were multiple recordings (Male01 and Female01). Within speaker analyses were carried out per speaker. The parameter time was divided into three time-stamps: 0–6 months, 6–12 months, and 12+ months for the male speaker, and 5 days and 11 days for the female speaker. Linear models were fitted to assess the change over time with the acoustic parameter as the dependant variable and Time and Voicing as the fixed effects. The model selection and criticism procedure used was identical to the one described above. For the fixed effects, post-hoc comparisons between the Time variable were performed using the *emmeans* function of version 1.7.2 of the *emmeans* package (Lenth, 2022).

The overall statistical significance of all post-hoc comparisons in this study was adjusted for false discovery rate at an alpha level of 0.05 in order to minimise the chances of reporting type-1 errors (Benjamini & Hochberg, 1995). The final model specifications can be

found in the Supplementary Materials (S-Table 3). All data and scripts pertaining to our acoustic measurements and statistical analyses can be accessed in our GitHub repository.[3]

4. Results

4.1. Research question (1) and (2): group analysis

The results of all hypothesis tests can be found in the Supplementary Materials (S-Table 4). Descriptive results (the mean and standard deviation) of the CoG and skewness parameter are displayed in Table 3, those of the BF parameter in Table 4, and the vowel formants in Table 5. Boldfaced cells indicate that the acoustic parameter was significantly different compared to the control group. When controlling for individual variation and time post-treatment, the clinical group produced all fricative groups with a significantly lower CoG as compared to the control group: labio-dental fricatives (est = -893.14 Hz, $p = 0.048$); dental fricatives (est = -882.44 Hz, $p = 0.048$); alveolar fricatives (est = -1870 Hz, $p = 0.0078$); post-alveolar fricatives (est = -1200 Hz, $p = 0.019$); affricates (est = −1520 Hz, $p = 0.0094$). Besides these main effects, there were three significant interactions between group and voicing such that the difference between clinical and typical speaker was larger when the phoneme was voiced: For alveolar fricatives (est = -1030 Hz, $p = 0.03$); post-alveolar (est = -1540 Hz, $p = 0.0078$); and for affricates (est = -842.32 Hz, $p = 0.05$). No other significant interactions were found.

As for the skewness parameter, the following fricative classes were significantly more positively skewed for clinical speakers: labio-dental fricatives (est = 2.04, $p = 0.045$); dental fricatives (est = 2.32, $p = 0.03$); and alveolar fricatives (est = 2.81, $p = 0.02$). There were no significant interactions between group and voicing for skewness.

For the statistical modelling of the BF parameter, we square-root-transformed the data in order to satisfy the assumption of normally distributed residuals. Affricates produced by clinical speakers had a significantly lower BF as compared to controls (est = -13.09 sqrt Hz, $p = 0.03$). Moreover, the interaction between group and voicing is significant, meaning that the difference between clinical and control speakers was 16.21 sqrt Hz larger in /dʒ/ as compared to /t/ /ʃ/ ($p = 0.03$). No other main effects or interactions were found for the BF parameter.

There were no significant group differences for the main effects of duration ($p = 0.3$) and spectral tilt ($p = 0.11$) nor for the F1 and F2 of the corner vowels.

4.2. Research question (3): longitudinal analysis

The results of all significance tests can be found in the Supplementary Materials (S-Tables 5–8). Results are shown based on how the acoustic parameters were at the later time frame as compared to the earlier time frame

4.2.1. Female01

To reiterate, T1 for Female01 denotes 5 days post-treatment and T2 11 days post-treatment. Descriptive results of the CoG and skewness parameter are displayed in Table 6, those of the BF parameter in Table 7, and the vowel formants in Table 8. Boldfaced cells indicate whether the acoustic parameter changed significantly over the course of the time frame. In terms of CoG, no significant differences were found. Note that this comparison is averaged over the voicing parameter as adding the interaction did not improve the model. For the skewness parameter, there was a significant difference for voiceless labio-dental fricatives only such that they were more positively skewed at T2 as compared to T1 (est = 0.55, $p = 0.006$). Other comparisons were not statistically significant.

As for the BF parameter, significant differences were found for /k/ and /dʒ/. For /k/, the BF was an estimated 506.6 Hz lower at T2 as compared to T1 ($p = 0.03$). For /dʒ/, the BF was an estimated 1411.1 Hz higher at T2 as compared to T1 ($p = 0.02$). All other comparisons did not reach significance.

Moving to the vowel formant measures, significant differences were found for the F1 of /i/ such that at T2, the F1 was an estimated 84.7 Hz higher than at T1 ($p < 0.001$). The F2 for /u/ was also significantly higher at T2 as compared to T1 (est = 440 Hz, $p = 0.01$). No other significant differences were found. Furthermore, there were no significant differences for the main effects of duration ($p = 0.55$) or spectral tilt ($p = 0.39$).

There were no significant main effects of duration and tilt. There were also no other significant results with regards to the BF, CoG, spectral skewness, or vowel formants.

4.2.2. Male01

For Male01, T1 are recordings 0 to 6 months post-treatment, T2 6 to 12 months post-treatment, and T3 12 months or longer post-treatment. Descriptive results of the CoG and skewness parameter are displayed in Table 9, those of the BF parameter in Table 10, and the vowel formants in Table 11. Again, the boldfaced cells indicate whether the acoustic parameter changed significantly over the time frame. T1-T2, T2-T3, and T1-T3 comparisons are discussed below.

4.2.2.1. T1-T2. At T2 as compared to T1, there was a main effect of spectral tilt such that a given phoneme had a significantly lower

[3] https://github.com/karkirowle/phoneme_paper

Table 3
Centre of Gravity and Skewness per phoneme per group.

Phoneme	Centre of Gravity in Hz (sd) Typical (sd)	Clinical	p-value	Skewness (sd) Typical	Clinical	p-value
/f/	2224 (469)	**1105 (585)**	**0.048**	0.74 (0.64)	**2.67 (1.58)**	**0.045**
/v/	983 (597)	**368 (203)**		2.82 (1.51)	**5.56 (2.24)**	
/θ/	2167 (712)	**1170 (657)**	**0.048**	0.90 (1.16)	**2.94 (1.76)**	**0.03**
/ð/	1178 (599)	**475 (350)**		2.63 (1.75)	**5.38 (2.03)**	
/s/	3812 (522)	**2678 (891)**	**0.0078***	-1.47 (1.01)	**0.04 (1.45)**	**0.02**
/z/	3398 (797)	**1109 (801)**		-1.03 (1.33)	**3.62 (2.44)**	
/ʃ/	3003 (606)	**2602 (659)**	**0.019***	0.02 (0.81)	0.18 (0.97)	0.2
/ʒ/	3222 (527)	**1123 (781)**		-0.57 (0.46)	3.00 (1.71)	
/tʃ/	3329 (555)	**2417 (696)**	**0.0094***	-0.31 (0.77)	0.62 (1.05)	0.053
/dʒ/	3130 (660)	**1325 (856)**		-0.23 (0.85)	2.75 (1.99)	

Note: the asterisk denotes a significant interaction between group and voicing. sd = standard deviation. The boldface indicates that the parameter was significantly different compared to the control group.

Table 4
Burst frequency per phoneme per group.

Phoneme	Burst frequency in Hz (sd) Typical	Clinical	p-value
/p/	1356 (838)	861 (700)	0.19
/b/	696 (579)	360 (301)	
/t/	2170 (1389)	2455 (1412)	0.71
/d/	1202 (1168)	703 (854)	
/k/	1720 (928)	1310 (925)	0.19
/g/	937 (749)	531 (536)	
/tʃ/	3604 (462)	**3267 (1082)**	**0.03***
/dʒ/	3343 (841)	**1602 (1480)**	

Note: the asterisk denotes a significant interaction between group and voicing. sd = standard deviation. The boldface indicates that the parameter was significantly different compared to the control group.

Table 5
Formant values of the corner vowels per group.

Vowel	Formant in Hz	Group Typical (sd)	Clinical (sd)	*p*-value
/i/	F1	435 (89)	370 (97)	0.52
	F2	2261 (283)	2132 (299)	0.38
/a/	F1	772 (101)	712 (147)	0.61
	F2	1278 (168)	1339 (282)	0.17
/u/	F1	453 (71)	390 (126)	0.60
	F2	1807 (336)	1661 (294)	0.13

Note: sd = standard deviation.

Table 6
Centre of Gravity and Skewness per phoneme per time frame for Female01.

Phoneme	Centre of Gravity in Hz (sd) T1	T2	p-value	Skewness (sd) T1	T2	p-value
/f/	2160 (539)	1531 (549)	**0.0502**	0.3 (1.0)	**1.4 (1.3)**	**0.006**
/v/	573 (486)	451 (201)		3.97 (2.0)	4.31 (1.9)	0.51
/θ/	1137 (425)	1515 (756)	0.45	2.82 (1.8)	1.8 (1.4)	0.46
/ð/	512 (392)	528 (343)		3.53 (1.3)	4.36 (1.5)	0.17
/s/	2820 (631)	2848 (738)	0.90	0.06 (1.2)	0.2 (1.0)	0.81
/z/	1165 (1270)	1469 (994)		5.6 (5.4)	2.91 (2.2)	0.33
/ʃ/	3171 (702)	2762 (105)	0.45	-0.46 (1.1)	0.47 (0.46)	0.54
/ʒ/	–	–		–	–	–
/tʃ/	–	2494 (907)	0.26	–	0.86 (1.1)	–
/dʒ/	704 (539)	1265 (783)		5.08 (3.8)	2.72 (1.8)	0.18

Note: the dashes indicate that the speaker did not produce this phoneme in this time frame. sd = standard deviation. The boldface indicates that the parameter changed significantly over the course of the time frame.

Table 7
Burst frequency per phoneme per time frame for Female01.

Phoneme	Burst frequency in Hz (sd) T1	T2	p-value
/p/	793 (258)	1148 (662)	0.25
/b/	392 (260)	569 (547)	0.54
/t/	2325 (1314)	2534 (1207)	0.08
/d/	366 (281)	602 (807)	0.82
/k/	1758 (831)	**1251 (659)**	**0.03**
/g/	803 (504)	769 (750)	0.93
/tʃ/	–	3200 (1309)	–
/dʒ/	240 (96)	**1651 (1325)**	**0.01**

Note: the dashes indicate that the speaker did not produce this phoneme in this time frame. sd = standard deviation. The boldface indicates that the parameter changed significantly over the course of the time frame.

Table 8
F1 and F2 per vowel per time frame for Female01.

Vowel	Formant in Hz	Time T1 (sd)	T2 (sd)	*p*-value
/i/	F1	403 (61)	**430 (60)**	**0.0028**
	F2	2012 (501)	2302 (242)	0.084
/a/	F1	718 (156)	663 (69)	0.36
	F2	1454 (461)	1349 (298)	0.98
/u/	F1	376 (92)	436 (74)	0.34
	F2	1573 (526)	**2014 (348)**	**0.012**

Note: sd = standard deviation. The boldface indicates that the parameter changed significantly over the course of the time frame.

Table 9
Centre of Gravity and Skewness per phoneme per time frame for Male01.

Phoneme	Centre of Gravity in Hz (sd) T1	T2	T3	p-values T1-T2	T2-T3	T1-T3	Skewness (sd) T1	T2	T3	p-values T1-T2	T2-T3	T1-T3
/f/	1023	858	1014	0.65	0.65	0.82	2.91	3.18	2.99	0.16	0.16	0.69
	(439)	(371)	(548)				(1.5)	(1.2)	(1.7)			
/v/	253	322	382				7.11	**5.39**	5.06	**<0.001**	0.21	**<0.001**
	(159)	(86.7)	(166)				(2.5)	**(1.7)**	(1.6)			
/θ/	1340	**914**	**1340**	**0.054**	**0.02**	0.22	2.78	**3.05**	**2.78**	**0.03**	**0.05**	0.65
	(711)	**(357)**	**(785)**				(2.1)	**(1.1)**	**(1.95)**			
/ð/	348	**420**	**470**				6.46	**4.97**	5.32	**<0.001**	0.36	**<0.001**
	(299)	**(107)**	**(322)**				(2.36)	**(1.23)**	(1.61)			
/s/	2643	**2062**	**2804**	**<0.001**	**<0.001**	**<0.001**	0.12	0.63	-0.05	0.88	0.85	0.85
	(625)	**(458)**	**(705)**				(1.03)	(0.8)	(0.9)			
/z/	845	**771**	**1324**				4.64	**3.46**	**2.46**	**<0.001**	**<0.001**	**<0.001**
	(708)	**(414)**	**(753)**				(2.6)	**(1.4)**	**(1.75)**			
/ʃ/	2554	**2130**	**2750**	**0.007**	**<0.001**	0.07	0.29	0.41	0.1 (0.7)	0.38	0.58	0.46
	(464)	**(303)**	**(536)**				(0.6)	(0.41)				
/ʒ/	1061	**654**	**1021**				2.98	3.09	3.38	0.97	0.97	0.97
	(507)		**(677)**				(1.4)		(2.1)			
/tʃ/	2563	**1670**	**2675**	**<0.001**	**<0.001**	0.065	0.3 (0.7)	**1.37**	**0.42**	**0.04**	**0.04**	0.9
	(457)	**(356)**	**(610)**					**(0.8)**	**(0.7)**			
/dʒ/	1022	**820**	**1294**				3.42	3.38	2.74	0.91	0.22	0.1
	(642)	**(444)**	**(732)**				(1.7)	(1.2)	(1.6)			

Note: the standard deviation could not be computed for sounds that were produced less than three times. sd = standard deviation. The boldface indicates that the parameter changed significantly over the course of the time frame.

spectral tilt (est = -7.16 dB/decade, $p = <0.001$). The duration of the phonemes did not significantly differ between T1 and T2 ($p = 0.37$).

For the CoG parameter, a significant decrease was observed at T2 for alveolar fricatives (est = -450 Hz, $p < 0.001$); post-alveolar fricatives (est = -415.7 Hz, $p = 0.007$); and affricates (est = -545.2 Hz, $p = 0.006$). Note that the results are averaged over voicing as the interaction with voicing did not improve the model fit.

For the skewness parameter, we square-root transformed the data and found a significant decrease in spectral skewness at T2 as compared to T1 for /v/ (est = -0.35, $p < 0.001$), /ð/ (est = -0.26, $p < 0.001$), /z/ (est = -0.3, $p < 0.001$). A significant increase in spectral skewness was found at T2 for /θ/ (est = 0.22, $p = 0.03$) and /tʃ/ (est = 0.35, $p = 0.04$).

For the BF parameter, we square-root transformed the data and found a significant decrease at T2 as compared to T1 was found for

Table 10
Burst frequency per phoneme per time frame for Male01.

	Burst frequency in Hz (sd)			p-values		
Phoneme	**T1**	**T2**	**T3**	**T1-T2**	**T2-T3**	**T1-T3**
/p/	724 (532)	566 (356)	689 (644)	0.98	0.98	0.98
/b/	245 (173)	319 (204)	272 (163)			
/t/	2474 (1509)	**1644 (1007)**	**2507 (1360)**	**<0.001**	**<0.001**	**0.03**
/d/	492 (689)	**487 (498)**	**794 (865)**			
/k/	1041 (660)	**694 (551)**	**1263 (944)**	**0.007**	**<0.001**	**0.008**
/g/	328 (250)	**338 (126)**	**479 (422)**			
/tʃ/	3469 (765)	**1637 (701)**	**1543 (1398)**	**0.005**	**<0.001**	0.17
/dʒ/	941 (944)	**835 (620)**				

Note: sd = standard deviation. The boldface indicates that the parameter changed significantly over the course of the time frame.

Table 11
F1 and F2 per vowel per time frame for Male01.

		Time			p-values		
Vowel	**Formant in Hz**	**T1 (sd)**	**T2 (sd)**	**T3 (sd)**	**T1-T2**	**T2-T3**	**T1-T3**
/i/	F1	269 (88)	**386 (58)**	**336 (62)**	**<0.001**	**<0.001**	**<0.001**
	F2	2016 (217)	**1957 (144)**	1961 (160)	**0.01**	0.8	**0.004**
/a/	F1	722 (202)	745 (11)	703 (123)	0.25	0.14	0.41
	F2	1339 (330)	1268 (232)	1328 (283)	0.52	0.52	0.24
/u/	F1	333 (177)	**401 (42)**	407 (121)	**<0.001**	0.2	**<0.001**
	F2	1685 (334)	1743 (211)	1632 (237)	0.28	0.2	0.28

Note: sd = standard deviation. The boldface indicates that the parameter changed significantly over the course of the time frame.

alveolars plosives (est = -5.5 sqrt-Hz, $p < 0.001$), velar plosives (est = -4.6 sqrt-Hz, $p = 0.007$), and affricates (est = -10.3 sqrt-Hz, $p = 0.005$).

For the vowel formants, we log-transformed the F1 and found a significant increase at T2 compared to T1 for /i/ (est = 0.4 log-Hz, $p < 0.001$), and /u/ (est = 0. 33 log-Hz, $p < 0.001$). For the F2, we found a significant decrease for /i/ (est = -67.2 Hz, p = 0.01).

4.2.2.2. T2-T3. At T3 as compared to T2, there was a main effect of spectral tilt such that a given phoneme had a significantly higher spectral tilt (est = 9.34 dB/decade, $p = <0.001$). There was also an effect of duration, for which we square-root transformed the data, such that speech became faster at T3 as compared to T2 as shown by a significant decrease in duration (est = -0.01 sqrt-s, $p = 0.02$).

For the CoG parameter, we observed a significant increase at T3 as compared to T2 for dental fricatives (est = 237.3 Hz, $p = 0.02$); alveolar fricatives (est = 717.6 Hz, $p < 0.001$); post-alveolar fricatives (est = 601.8 Hz, $p < 0.001$); and affricates (est = 737.6 Hz, $p < 0.001$).

For the skewness parameter, the data was square-root-transformed. At T3, we observed a significant decrease in spectral skewness for /θ/ (est = -0.19, $p = 0.05$); /z/ (est = -0.34, $p < 0.001$); and /tʃ/ (est = -0.36, $p = 0.04$).

For the BF parameter, the data was square-root-transformed. A significant increase in BF was found at T3 as compared to T2 for alveolar plosives (est = 7.6 sqrt-Hz, $p < 0.001$); velar plosives (est = 7.5 sqrt-Hz, d$p < 0.001$); and affricates (est = 13.7 sqrt-Hz, $p < 0.001$).

Lastly, for the vowel formants, we observed a significant decrease in F1 only for /i/, which was log-transformed (est = -0.14 log-Hz, $p < 0.001$).

4.2.2.3. T1-T3. Finally, we compared T3 against T1 to investigate whether speech continued to improve 12 months or longer after treatment. At T3, there was a main effect of spectral tilt such that a given phoneme had a significantly higher spectral tilt (est = 2.18 dB/decade, $p = 0.02$). There was also a main effect of duration as phonemes became significantly shorter at T3 (est = -0.008 sqrt-s, $p = 0.02$).

For the CoG parameter, alveolar fricatives had a significantly higher CoG compared to T1 (est = 267.4 Hz, $p = <0.001$). No other significant effects were found.

For the skewness parameter, for which the data was square-root transformed, we found a significant decrease in skewness for /v/ (est = -0.44, $p <0.001$); /ð/ (est = -0.19, $p < 0.001$); and /z/ (est = -0.63, $p < 0.001$). No other significant effects were found.

As for BF, for which we square-root transformed the data, we found significant differences between alveolar plosives, such that the BF was higher at T3 compared to T1 (est = 2.06 sqrt-Hz, $p = 0.03$); velar plosives (est = 2.93 sqrt-Hz, $p = 0.008$). No other significant effects were found.

Lastly, we found significantly higher F1 values for /i/ (est = 0.27 log-Hz, $p < 0.001$) and /u/ (est = 0.28 log-Hz, $p < 0.001$). The F2 of /i/ was also significantly lower at T3 compared to T1 (est = -60 Hz, $p = 0.004$). No other significant differences were found.

5. Discussion

5.1. Group analysis

The aim of the present study was to answer three questions concerning the impact of oral cancer surgery on the acoustics of the speech signal. The first two questions concern how key acoustic features change in the speech of individuals who were surgically treated for oral cancer and whether these changes were related to specific PoAs or MoAs. This question was addressed from a large-scale acoustic comparison of clinical and typical spontaneous speech. Subsections of MoA and PoA were analysed on various acoustic parameters. It was predicted that (1) fricatives would be distorted, especially [+] coronal sounds; (2) plosives would be distorted; and (3) that there would be more vowel centralisation post-treatment.

Even though speaking slower is a common compensatory mechanism in this population we did not find any significant differences in terms of duration of the phoneme (Bressmann et al., 2010; Constantinescu & Rieger, 2019). However, an informal comparison in terms of speech rate does suggest a slight difference (2.95 syllables/second in the clinical group compared to 3.51 syllables/second in the control group). This could be due to the automatic nature of the speech rate script whereas the phonemes were manually segmented. Likewise, we did not find any differences in terms of relative loudness, a parameter ITOC may use to compensate for deviant speech.

Group differences were found in fricatives and affricates (see Table 3). In these classes, the speech of ITOC had a lower Centre of Gravity (CoG) as compared to typical speech. These results confirm our first prediction and are in line with previous research that found group differences in the production of fricatives (e.g., Acher et al., 2014; Jacobi et al., 2013; Laaksonen et al., 2011; Zhou et al., 2011). The present study adds labio-dental fricatives, a class that has not been analysed before. In general, a lower CoG is associated with a more backed pronunciation of the fricative due to a more posterior constriction of the tongue, resulting in a smaller cavity length posterior of the constriction (Zhou et al., 2011). This more backed pronunciation may be due to restricted tongue movements, either as a result of surgical treatment or due to CRT (e.g., Bressmann et al., 2004). Moreover, a lower CoG might be indicative of a reduced ability (e.g., in terms of muscle strength or precision) to sustain the high pressure resulting from the narrow and precise constriction found in sibilants. Our results, however, diverge from those by Laaksonen et al. (2011) regarding post-alveolar fricatives. They reported effects on alveolar fricatives, but not on post-alveolars whereas we found differences in both classes. This discrepancy may be interpreted in two ways. First, our speakers may have more restricted tongue movements than those in Laaksonen et al. (2011) which may cause the differences in both fricative classes. Alternatively, post-alveolars may be produced with a lower CoG in order to maintain acoustic differentiation between /s/ and /ʃ/, thus exhibiting some form of compensatory behaviour. However, this second explanation is less probable as Table 3 indicates that /s/ and /ʃ/ are produced with almost identical CoG values, which makes an explanation based on reduced mobility more likely.

It is worth noting that significant differences were found between all fricative classes despite the fact that all but one recording (Female01 recording 1) had a perceptual rating above 4, indicating that speech was easy to understand and transcribe. This finding shows that clinical speakers may be fully intelligible despite having atypical acoustic productions. Speech in the current study came from recordings without background noise or music, creating an ideal listening environment. Eadie et al. (2021) reported that the intelligibility of ITOC who exhibit mild speech imprecisions might be especially vulnerable to effects of increased background noise as these environments drastically reduced the intelligibility of speech. Future work might consider the relationship between degree of acoustic deviance and loss of intelligibility in noise as this might show the clinical relevance of acoustic measures even more.

Turning to the plosive results, the data do not support our second prediction that this class would be distorted. Although many previous studies (e.g., Bressmann et al., 2004; De Bruijn et al., 2009; Halpern et al., 2020; Jacobi et al., 2013) found that plosives pose difficulty for ITOC, the data presented here indicate that the burst frequency (BF) at all PoAs produced by ITOC showed no significant difference when compared to typical speech (see Table 4). Only affricates were produced with a lower BF. However, it is evident from Table 4 that the BF tends to be lower in the clinical group, albeit non-significantly lower. Although a clear explanation is lacking, some suggestions can be made. First, approximately half of our corpus (16.279 tokens) is from speech from one year or longer post-treatment and from speakers who were intelligible at the word level. It could be the case that by this time, plosives are produced with a typical BF or due to the relatively high base level of intelligibility as evidenced by the SLP ratings. This would go against the findings of De Bruin et al. (2009) and Jacobi et al. (2013) as they both reported significant group differences for plosives one year post-treatment. Although Jacobi et al. (2013) compared pre-treatment to post-treatment recordings whereas control speech was used in the present study, the difference still stands. This is because the pre-treatment baseline could already be perturbed by the presence of the tumour and was found to further deteriorate after treatment whereas we found no differences between ITOC and controls. Second, provided that there is considerable individual variation in the speech outcomes following oral cancer treatment (Bressmann, 2013), it is plausible that plosives did not pose as much difficulty for our group of speakers. Including a larger sample speaker wise one year post-treatment would shed light on this possibility. Third, the size of our dataset might explain the lack of significant differences. In previous acoustic studies (e.g., De Bruin et al., 2009), a limited number of tokens per phoneme were used in the analysis whereas in the present study we used more than 30 tokens per phoneme for most speakers, for Male01 even over 150. This larger dataset helps to masque the large individual variation that is typically found in speech which raises the question of whether building up and releasing the right amount of pressure pose as much difficulty as was previously assumed. However, more research with larger sample sizes token-wise is needed in order to verify this suggestion.

The data do also not support our third prediction that vowel centralisation would increase post-treatment (see Table 6). Some studies found centralisation of the VSA even one year post-treatment (e.g. Jacobi et al., 2013; Takatsu et al., 2017). Others, on the other hand, found only durational and fundamental frequency differences one year post-treatment (cf. Laaksonen et al., 2010). The data

presented here seems to support the findings by Laaksonen et al. (2010) in that there were no significant differences in the first and second formants between clinical and typical speech (see Table 5). As noted before, the speech corpus primarily consists of speech after this one year period which makes it plausible that the included speakers have regained the ability to produce the corner vowels with typical formant values by this time. We did observe relatively high F2 values for /u/. Even though a high F2 for /u/ is expected given the ongoing u-fronting sound change (e.g., Clopper et al., 2019; Hillenbrand et al., 1995), it seems more likely that the third formant (F3) was mistracked as the F2. Therefore, the results need to be interpreted with caution and future studies with large token numbers need to be more careful with automatic formant extraction, for example by using custom formant ceilings per speaker and vowel (see e.g., Escudero et al., 2009).

For voicing, the data suggest an effect of voicing on (post-) alveolar fricatives and affricates as evidenced by significant interactions between group and voicing such that the voiced variants had poorer acoustic outputs. Although we did not formulate precise predictions, we would expect that the locus of the communication problems were in the articulatory domain as opposed to the voicing domain since the voice box is spared in surgical treatment for oral cancer. However, some of the clinical speakers received CRT, namely Male01, Female02, and Female03, which is known to affect the quality of speech due to a stiffening of the tongue (e.g., Jacobi et al., 2013; Lazarus, 2009). If the CRT was employed in laryngeal regions, for example to radiate submandibular or anterior cervical lymph nodes, problems may arise with hydrating the vocal folds, which could result in problems with voicing (Lazarus, 2009; Sato and Nakashima, 2008). Male01 received CRT targeting the lymph nodes, but the recording does not specify which specific ones, making it impossible to verify this suggestion in the present study and leaving it for future inquiries.

The second research question concerned whether changes in articulatory precision were related to specific manners or places of articulation. To assess this, the data was analysed by making different PoA and MoA subsections. It was predicted that PoAs that do not directly involve the tongue in its production (e.g., bilabials) would pose less difficulty for clinical speakers than PoAs that do actively use the tongue (e.g., alveolars). This prediction was not borne out by the data. On the one hand, the results show that neither bilabial nor alveolar plosives were significantly affected by surgical oral cancer treatment. On the other hand, all classes of fricatives were affected and these distortions are found over a range of PoAs, including those that do not involve the tongue (see Table 3). Both [-] coronal fricatives (e.g., labio-dental fricatives) and [+] coronal fricatives (e.g., (post)alveolar fricatives) were affected. Paired with the absence of an effect for plosives, the results support the notion that it is MoA that is more relevant in identifying problem areas for individuals treated for oral cancer than PoA. Furthermore, this finding would point to general problems with articulatory precision as fricatives require narrow constrictions through which a turbulent airflow escapes. Reduced precision is most likely a side-effect of the reduced tongue mobility, or reduced control, caused by the surgical procedure. For example, Zhou et al. (2013) provided cine-MRI data that showed that the transition in /i.ʃi/ was more problematic than the transition in /a.ʃa/ due to the similar tongue shapes and place of articulation needed for both /i/ and /ʃ/. Moreover, there is a remote possibility that the facial nerves innervating the lips could have been affected by the treatment which would explain the differences in labiodentals, but this possibility could not be verified. Future research could further explore the articulatory precision by (1) analysing the acoustics or kinematics of precise consonant-vowel transitions using multiple fricatives and plosives; and (2) investigating whether general lip movement is impaired or not due to damage to the facial nerves. If the same effect is found for plosives as well, this would provide stronger evidence in favour of general problems with articulatory precision. Still, we must keep in mind that these conclusions are based on the target PoA and MoA. As acoustics can only provide indirect evidence of the PoA, more articulatory data from this population is needed to verify our results.

We have hypothesised that ITOC might have difficulty with articulatory precision as the treatment might limit the mobility of and control over the articulators. If this is indeed the case, one would predict more acoustic variability in the output. This does not seem to be the case consistently as standard deviations are not higher in the clinical group. However, standard deviations are higher for alveolar and post-alveolar fricatives in the clinical group, which require a very complex and precise constriction. Although this finding seems to support our precision hypothesis, results are not consistent in plosives and vowels and we have to keep in mind that the standard deviations only give an impression of the variability as we did not conduct formal tests. Future work might consider the variability from a dynamic viewpoint by looking at the entire segment (e.g., formant trajectories) instead of measuring the acoustic parameters at a single time point. Following our articulatory precision hypothesis, we would predict less stable phone segments that contain more variability.

5.2. Longitudinal analysis

Our third question targeted the temporal change of the acoustic parameters. This question was addressed by comparing clinical speech at different time frames. It was predicted for Male01 that (1) vowel centralisation should improve post-treatment (i.e., less centralisation); (2) the BF of plosives should increase; and (3) the CoG of fricatives should increase while their skewness should decrease. For Female01 we refrained from formulating any predictions due to the timing of the two post-treatment recordings.

The results of Female01, for whom we compared recordings of 11 days post-treatment to 5 days post-treatment, indicate that there was significant vowel centralisation due to an increase of F1 of /i/ (lower tongue position) and an increase of F2 of /u/ (more anterior position). No changes were observed for /a/. This is partly in line with previous research that found a centralisation of the VSA 14-days post-treatment (Takatsu et al., 2017). However, Takatsu et al. (2017) based their conclusion on a comparison of the individual's speech pre- and post-treatment whereas we compared two post-treatment recordings. Our results add to Takatsu et al. (2017) that the vowels may become more centralised during the first weeks of the recovery process which could be due to the pain and swelling associated with the recovery. Another possibility is that the patient is still exploring compensatory strategies and has not yet found a satisfactory strategy that can produce better quality acoustic output.

As for plosives, the BF of /k/ decreased at 11 days post-treatment as compared to 5 days post-treatment. This finding highlights the

speaker specific variation since we did not detect BF differences on the group level. As with vowel centralisation, the swelling and pain associated with the surgical recovery may cause problems in producing /k/ and /g/, although the difference in the case of /g/ was non significant. We found that the BF of /dʒ/ did increase at T2 (1651 Hz) as compared to T1 (240 Hz). Even though it is common that voiced affricates have a lower BF than their voiceless counterparts, we note that the BF of /dʒ/ was exceptionally low at T1 (Chodroff and Wilson, 2014). If this is not caused by a measurement error, we can conclude that (post)-alveolar plosives improve at a faster rate than velar plosives. Discrepant improvement could result from the place of the resection (i.e., the wound) or a difference in hinder caused by the swelling and pain. However since we do not have detailed information regarding the speaker's surgery, this finding needs to be replicated in future inquiries to gain more credibility.

Lastly, the lack of improvement found in fricatives is hypothesised to be due to the time frame of the recordings. This class of sounds probably needs more time to improve as fricatives are characterised by narrow and precise constrictions which require very sophisticated spatial coordination. This interpretation is backed by the results of Laaksonen et al. (2011), who reported that speakers only showed major improvements at the six-months mark while speech became more distorted at the one-month mark. The only significant observed difference was an increase in spectral skewness for /f/ which denotes that more energy was localised on the lower frequency bands.

Turning to Male01, we compared three time points: T1 as 0–6 months, T2 as 6–12 months, and T3 as more than 12 months post-treatment. Prediction (1), that there would be less vowel centralisation, is not supported by the data. On the contrary, we observed a more vowel centralisation at T2, most likely due to reduced tongue mobility. The F1 was higher for /i/ and /u/, denoting reduced tongue height. Furthermore, we observed a reduced F2 for /i/, denoting reduced anterior lingual movement. Results are not in line with the previous literature that indicated improvement in the VSA after an on average 50-day rehabilitation period (Takatsu et al., 2017). Results are more in line with Laaksonen et al. (2010) who observed an increase in F1 post-treatment for male ITOC. The reduced vowel contrastivity was partially made up for at T3 as the F1 of /i/ became significantly lower compared to T2, indicating increased tongue height in the production of /i/. However, the F1 of /u/ and F2 of /i/ did not improve significantly. Our results therefore show some long-lasting changes in terms of vowel formants.

In general, Male01 had an interesting trajectory. Speech was found to become acoustically more distorted at T2 as compared to T1. A possible explanation could be a fatigue effect, resulting in speech becoming more centralised. This is evidenced by the centralisation of the vowels /i/ and /u/, as well as a decreased CoG in dental fricatives, (post) alveolar fricatives, affricates, and a decrease in BF for affricates, alveolar plosives, and velar plosives. These observations together with the fact that there was only one recording from this speaker at T2, makes it more plausible that extralinguistic factors like fatigue might have influenced the acoustic quality. However, at T2, Male01 spoke louder as shown by the decrease in spectral tilt which would not be expected if the changes are due to fatigue. An alternative explanation is that the relative loudness of his speech was used as a compensatory mechanism in order to maximise intelligibility and compensate for the acoustic reduction. Regardless of the correct explanation, the data are not in line with predictions (2) and (3) at T2 for Male01.

This unexpected decrease in acoustic quality was made up for at T3. The CoG of all affected fricatives and affricates increased, and the BF of affricates, and alveolar and velar plosives significantly increased, as well. For alveolar fricatives, and alveolar and velar plosives, these improvements exceeded even the T1 measures, signalling long term improvement. These results are in line with predictions (2), that BF of plosives should increase and (3) that the CoG of fricatives should increase. Furthermore, Male01 spoke significantly faster at T3 and the spectral tilt increased significantly, signalling that speech had become more quiet again. These last two changes might be a result of the acoustic improvement found at T3, such that the intended compensation strategy for acoustic reduction at T2 in terms of relative loudness was not necessary anymore after one year. However, it should also be noted that the /i/ was still more centralised at T3 compared to T1, showing that not all sounds improved over time.

Although the present study has presented novel results, the study was not without limitations. First, the present study had a lack of phonetic control. Coarticulation and lexical stress were not controlled for in the consonant analysis. Stress was not a limitation for the vowel analysis as we did not include /ə/ realisations, but the limitation of coarticulation still persists. However, we tried to account for coarticulation as much as possible by measuring the formant at the midpoint of the vowel rather than taking the mean formant value over the entire vowel segment. The sheer size of the dataset further masks some of this coarticulation induced variation. Second, we observed mistracking of the F2 of /u/ in our data due to the automatic formant extraction. Future studies could use speaker or vowel specific formant ceilings in order to improve on this aspect. Third, only five speakers were included in the analysis despite the fact that speech outcomes are highly variable (Bressmann, 2013). Yet, clear differences were found for fricatives, showing the robustness of the effect despite speaker variation. Fourth, we only had multiple recordings for two speakers which reduced our ability to provide stronger generalisations for our longitudinal analysis. Fifth, our analysis included individuals who were treated for tongue or jaw tumours, which might masque effects if their data are pooled together. However, kinematic investigations using electromagnetic articulography suggest that ITOC who received a (partial) jaw replacement exhibit reduced tongue mobility as well (Tienkamp et al., 2022). We therefore decided not to split our data, yet it could be possible that certain effects were masked. Sixth, online platforms like YouTube provide us with an unprecedented opportunity to collect large datasets. However, this comes at the expense of rigorous control that is often preferred in phonetic research. One cannot control for the time point post-treatment, microphone quality, and environmental noise of the recording. For example, one recording was excluded due to having background music. With the inability to control for time points, it becomes rather difficult to collect pre- and post-treatment speech of the same speaker, which, in turn, makes it hard to draw firm conclusions on the temporal development. Seventh, due to our selection criterion of being intelligible at the word level, we introduced a form of speaker bias which is also shown by the SLP ratings, which were relatively high. This bias might have resulted in the absence of any plosive and vowel effects. Future work with target sentences or words could remedy this problem as participants had to be intelligible in order to transcribe the unprompted speech in the present study. Lastly, we did not have

diachronically matched control speakers for the ITOC's speech. By using control corpora like TIMIT, it was impossible to directly match the diachronic variety of English for all speakers. However, this would have been a limitation for all available control corpora. We therefore opted for the TIMIT corpus as phonetic transcriptions were already available. Still, it can thus not be guaranteed that acoustic shifts did not have an impact on the results, but we actively tried to match controls to ITOC based on their diachronic variety of English as best as possible.

6. Conclusion

This paper presented the first acoustic analysis of the spontaneous speech of individuals who have been surgically treated for oral cancer. This analysis was done by using a spontaneous speech corpus containing one hour of speech. All sounds with a fricative manner of articulation were found to be distorted post-treatment which shows that manner of articulation is more predictive of articulatory changes post-treatment than place of articulation. Crucially, both [-] and [+] coronal fricatives were distorted. Moreover, a possible effect of chemoradiation therapy was found such that it negatively affected the production of voiced (post-) alveolar fricatives and affricates. No evidence was found for the improvement of the VSA, most likely because improvement happens within the first 50 days whereas the corpus primarily consists of recordings longer than one year post-treatment. The within speaker analysis showed that the centre of gravity of fricatives, as well as the burst frequency of plosives increase over time, but later than has previously been reported, namely after the one-year mark.

Future research could investigate the articulatory precision of ITOC further by comparing and contrasting dynamic aspects of phonemes (e.g., formant trajectories) or consonant vowel transitions and coarticulatory behaviour that have similar or different places of articulation (e.g., /ti/ and /tu/ versus /ki/ and /ku/). If the articulatory precision is mostly affected, one would predict more difficulty with /ti/ and /ku/ as they partly share the place of articulation. Another direction would be to investigate whether the temporal development of consonants is related to the site of the resection. If this is the case, one would predict that, if a patient lost more of the anterior part of the tongue, alveolar consonants would take longer to improve as compared to velar consonants. Lastly, a direct comparison between spontaneous speech and read-speech might be a fruitful direction in order to probe the usefulness of spontaneous speech in clinical phonetics since many of the initial assumptions were not supported by the data presented here.

Ethics

Ethical approval was not sought because all data used in the present study is publicly available on YouTube and previously published (Halpern et al., 2020).

CRediT authorship contribution statement

Thomas B. Tienkamp: Conceptualization, Methodology, Data curation, Formal analysis, Writing – original draft, Visualization, Writing – review & editing. **Rob J.J.H. van Son:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – review & editing, Supervision, Project administration. **Bence Mark Halpern:** Conceptualization, Methodology, Investigation, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest

Acknowledgments

The Department of Head and Neck Oncology and Surgery of the Netherlands Cancer Institute receives a research grant from Atos Medical (Hörby, Sweden), which contributes to the existing infrastructure for quality of life research. The funders had no active role in the design, execution, or analysis of the research. The authors want to thank the editor and two reviewers whose comments improved the quality of our manuscript. Any remaining mistakes are obviously our own.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jcomdis.2022.106292.

Appendix A

Instructions used for the SLP ratings.
Instruction reminder
A "healthy" utterance is a 5, by which we mean:

- The utterance is easy to understand.

- You could wine down the utterance, if asked, without any difficulties.
- The speaker has a clear articulation with a normal speaking rate.

A severely "pathological" utterance is a 1, by which we mean:

- The utterance is impossible to uanscribe. even after repeated listening attempts.
- The speaker has articulation problems.
- The speaker might speak slower or faster than normal.

References

Acher, A., Perrier, P., Savariaux, C., & Fougeron, C. (2014). Speech production after glossectomy: Methodological aspects. *Clinical Linguistics & Phonetics, 28*(4), 241–256. https://doi.org/10.3109/02699206.2013.802015

Baayen, R. (2008). Analyzing Linguistic Data: A Practical Introduction to Statistics using R. Cambridge: Cambridge University Press. 10.1017/CBO9780511801686.

Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences*. Palgrave Macmillan.

Balaguer, M., Pommée, T., Farinas, J., Pinquier, J., Woisard, V., & Speyer, R. (2020). Effects of oral and oropharyngeal cancer on speech intelligibility using acoustic analysis: Systematic review. *Head & Neck, 42*(1), 111–130. https://doi.org/10.1002/hed.25949

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

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Blyth, K. M., McCabe, P., Madill, C., & Ballard, K. J. (2015). Speech and swallow rehabilitation following partial glossectomy: A systematic review. *International journal of speech-language pathology, 17*(4), 401–410. https://doi.org/10.3109/17549507.2014.979880

Boersma, P., & Weenink, D. (2020). *Praat: Doing Phonetics by computer [computer programme]* (6.20.06). https://praat.org.

Bressmann, T. (2013). Head and neck cancer and communicationL. Cummings (Ed.), The Cambridge Handbook of Communication Disorders. *Cambridge University Press*, 161–184. https://doi.org/10.1017/CBO9781139108683.013

Bressmann, T., Jacobs, H., Quintero, J., & Irish, J. C. (2010). Speech outcomes for partial glossectomy surgery: Measures of speech articulation and listener perception. *Canadian Journal of Speech-Language Pathology and Audiology, 33*(4), 204–210.

Bressmann, T., Sader, R., Whitehill, T. L., & Samman, N. (2004). Consonant intelligibility and tongue motility in patients with partial glossectomy. *Journal of Oral and Maxillofacial Surgery, 62*(3), 298–303. https://doi.org/10.1016/j.joms.2003.04.017

Chodroff, E., & Wilson, C. (2014). Burst spectrum as a cue for the stop voicing contrast in American English. *The Journal of the Acoustical Society of America, 136*(5), 2762–2772. https://doi.org/10.1121/1.4896470

Clopper, C. G., Burdin, R. S., & Turnbull, R. (2019). Variation in /u/fronting in the American Midwest. *The Journal of the Acoustical Society of America, 146*(1), 233–244. https://doi.org/10.1121/1.5116131

Constantinescu, G., & Rieger, J.M. (2019). Speech Deficits Associated with Oral and Oropharyngeal Carcinomas. In P. C. Doyle (Ed.), *Clinical care and rehabilitation in head and neck cancer* (pp. 265–279). Springer International Publishing. https://doi.org/10.1007/978-3-030-04702-3_16.

De Bruijn, M. J., ten Bosch, L., Kuik, D. J., Quené, H., Langendijk, J. A., Leemans, C. R., & Verdonck-de Leeuw, I. M. (2009). Objective Acoustic-Phonetic Speech Analysis in Patients Treated for Oral or Oropharyngeal Cancer. *Folia Phoniatrica et Logopaedica, 61*(3), 180–187. https://doi.org/10.1159/000219953

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods, 41*(2), 385–390. https://doi.org/10.3758/BRM.41.2.385

Eadie, T. L., Durr, H., Sauder, C., Nagle, K., Kapsner-Smith, M., & Spencer, K. A. (2021). Effect of Noise on Speech Intelligibility and Perceived Listening Effort in Head and Neck Cancer. *American Journal of Speech-Language Pathology, 30*(3S), 1329–1342. https://doi.org/10.1044/2020_AJSLP-20-00149

ELAN. (2020). (Version 5.4) [Computer software]. Nijmegen: Max Planck Institute for Psycholinguistics. https://archive.mpi.nl/tla/elan.

Escudero, P., Boersma, P., Rauber, A. S., & Bion, R. A. (2009). A cross-dialect acoustic description of vowels: Brazilian and European Portuguese. *The Journal of the Acoustical Society of America, 126*(3), 1379–1393. https://doi.org/10.1121/1.3180321

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression (Third)*. Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., & Zue, V. (1993). Timit acoustic phonetic continuous speech corpus [LDC93S1]. *Linguistic Data Consortium*.

Ghio, A., Révis, J., Merienne, S., & Giovanni, A. (2013). Top-Down Mechanisms in Dysphonia Perception: The Need for Blind Tests. *Journal of Voice, 27*(4), 481–485. https://doi.org/10.1016/j.jvoice.2013.03.015

Hagedorn, C., Lu, Y., Toutios, A., Sinha, U., Goldstein, L., & Narayanan, S. (2022). Variation in compensatory strategies as a function of target constriction degree in post-glossectomy speech. *JASA Express Letters, 2*(4), Article 045205. https://doi.org/10.1121/10.0009897

Halpern, B.M., van Son, R.J.J.H., van den Brekel, M.W.M., & Scharenborg, O. (2020). Detecting and Analysing Spontaneous Oral Cancer Speech in the Wild. *Proc. Interspeech* 2020, 4826–4830. https://doi.org/10.21437/Interspeech.2020-1598.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America, 97*(5), 3099–3111. https://doi.org/10.1121/1.411872

Jacobi, I., van Rossum, M. A., van der Molen, L., Hilgers, F. J. M., & van den Brekel, M. W. M. (2013). Acoustic Analysis of Changes in Articulation Proficiency in Patients with Advanced Head and Neck Cancer Treated with Chemoradiotherapy. *Annals of Otology, Rhinology & Laryngology, 122*(12), 754–762. https://doi.org/10.1177/000348941312201205

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language, 45*, 326–347.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, (13), 82. https://doi.org/10.18637/jss.v082.i13

Laaksonen, J.-. P., Rieger, J., Happonen, R.-. P., Harris, J., & Seikaly, H. (2010). Speech after radial forearm free flap reconstruction of the tongue: A longitudinal acoustic study of vowel and diphthong sounds. *Clinical Linguistics & Phonetics, 24*(1), 41–54. https://doi.org/10.3109/02699200903340758

Laaksonen, J.-. P., Rieger, J., Harris, J., & Seikaly, H. (2011). A longitudinal acoustic study of the effects of the radial forearm free flap reconstruction on sibilants produced by tongue cancer patients. *Clinical Linguistics & Phonetics, 25*(4), 253–264. https://doi.org/10.3109/02699206.2010.525681

Lachman, M.E. (2001). Adult development, psychology of. In N. J.Smelser & P. B.Baltes (Eds.). *International encyclopaedia of the social & behavioral sciences* (pp. 135–139). https://doi.org/10.1016/b0-08-043076-7/01650-8.

Lazarus, C.L. (2009). Effects of chemoradiotherapy on voice and swallowing: *Current Opinion in Otolaryngology & Head and Neck Surgery*, 17(3), 172–178. https://doi.org/10.1097/MOO.0b013e32832af12f.

Lenth, R.V. (2022). *Emmeans: Estimated Marginal Means, aka Least-Squares Means*. https://CRAN.Rproject.org/package=emmeans.

Nicoletti, G., Soutar, D.S., Jackson, M.S., Wrench, A.A., Robertson, G., & Robertson, C. (2004). Objective Assessment of Speech after Surgical Treatment for Oral Cancer: Experience from 196 Selected Cases: *Plastic and Reconstructive Surgery*, 113(1), 114–125. https://doi.org/10.1097/01.PRS.0000095937.45812.84.

Oates, J. (2009). Auditory-Perceptual Evaluation of Disordered Voice Quality. *Folia Phoniatrica et Logopaedica, 61*(1), 49–56. https://doi.org/10.1159/000200768

Powell, M. J. (2009). *The bobyqa algorithm for bound constrained optimization without derivatives. report no. DAMTP, 2009/NA06* (p. 26). Centre for Mathematical Sciences, University of Cambridge.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Sato, K., & Nakashima, T. (2008). Effect of Irradiation on the Human Laryngeal Glands. *Annals of Otology, Rhinology & Laryngology, 117*(10), 734–739. https://doi.org/10.1177/000348940811701005

Takatsu, J., Hanai, N., Suzuki, H., Yoshida, M., Tanaka, Y., Tanaka, S., Hasegawa, Y., & Yamamoto, M. (2017). Phonologic and Acoustic Analysis of Speech Following Glossectomy and the Effect of Rehabilitation on Speech Outcomes. *Journal of Oral and Maxillofacial Surgery, 75*(7), 1530–1541. https://doi.org/10.1016/j.joms.2016.12.004

Tienkamp, T. B., Rebernik, T., Halpern, B. M., Abur, D., van Son, R. J. J. H., de Visscher, S. A. H. J., Witjes, M. J. H., & Wieling, M. (2022). Quantifying changes in articulatory working space following oral cancer treatment. In *Poster presented at the 8th International Conference on Speech Motor Control, Groningen, The Netherlands*.

van Sluis, K., Kapitein, M., van Son, R. J. J. H., & Boersma, P. (2019). The acoustic contrast between the Dutch consonants /t/and /d/ is reduced in tracheoesophageal speech. In *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia* (pp. 914–918), 2019: ICPhS2019: 5-9 August 2019.

Zhou, X., Stone, M., & Espy-Wilson, C. Y. (2011). A comparative acoustic study on speech of glossectomy patients and normal subjects. In *Proceedings of Interspeech 2011: 12th Annual Conference of the International Speech Communication Association* (pp. 517–520).

Zhou, X., Woo, J., Stone, M., & Espy-Wilson, C. (2013). A cine MRI-based study of sibilant fricatives production in post-glossectomy speakers. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 7780–7784). https://doi.org/10.1109/ICASSP.2013.6639178