UNIVERSITAT DE BARCELONA

Estudios de Fonética Experimental
Journal of Experimental Phonetics

LF Laboratori de Fonètica

Phonetic characteristics of spontaneous speech in a total laryngectomized Italian speaker: Perspectives for speech enhancement algorithms

Chiara Meluzzi[a], Sonia Cenceschi[b], Francesco Roberto Dani[c], Alessandro Trivilini[d]

[a] University of Milan "La Statale" (Italy), chiara.meluzzi@unimi.it
[b] University of Applied Sciences and Arts of Southern Switzerland (Switzerland), sonia.cenceschi@supsi.ch
[c] University of Applied Sciences and Arts of Southern Switzerland (Switzerland), francesco.dani@supsi.ch
[d] University of Applied Sciences and Arts of Southern Switzerland (Switzerland), alessandro.trivilini@supsi.ch

ARTICLE INFO

Article history
Received: 18/02/2021
Accepted: 13/03/2022

Keywords
Oesophageal Speech
Voice quality
Voice disorders
Vowel formants
Clinical phonetics

ABSTRACT

This paper describes the main phonetic features of an Italian L1 74 y. o. speaker (ESO01) after he endured total laryngectomy in 2015 with the complete removal of vocal folds due to five tumour masses. We offer an acoustic analysis of the spontaneous speech of this target speaker, in order to lay ground to the development of spontaneous speech enhancement and reconstruction algorithms for non-invasive aids. A semi-automatic analysis extracts formants' values (F0, F1, F2, F3) on the midpoint and on 7 time-points, together with other acoustic cues. Our results show that our target speaker presents a low and rough voice, but his vowels are clearly differentiated. Furthermore, we find vocoid and air release to be extremely consistent in his acoustic characteristics during oesophageal phonation.

1. Introduction

This paper describes the main phonetic features of a 74-year-old Italian L1 speaker after he endured total laryngectomy five years prior to the recording, in 2015. Our speaker has suffered from neck cancer with five different metastasis sites all around the vocal cords. The surgeons performed a radical neck dissection with the complete removal of the vocal folds and the adjacent muscles (cf. Cummings & Cooper, 2008). Bressmann (2010, p. 503) has summarized that after a total laryngectomy, a permanent tracheotomy in the patient's lower neck, near the sternum, can assure patient's ventilation of the lungs; the patient now breaths exclusively through the neck. After this intervention, our speaker has learnt to generate voice as oesophageal speech (henceforth, ES; cf. Doyle & Finchmen, 2009). Previous works have focused on phonetic features in laryngectomized speakers and on their intelligibility across methods of alaryngeal speech (e.g., Williams & Watson, 1987). However, the vast majority of the work has been conducted on elicited speech or on sustained vowel production, in particular /a/, in order to check the impact of speech therapy after surgery.

Conversely, in this work we focus on segmental features as produced by our target speaker during a reading of list of sentences and a long spontaneous conversation on various topics. This study aims at providing a preliminary acoustic description of Italian ES in naturally occurring conversation. Based on these findings, we aim at laying the foundations for future algorithmic implementations focused on improving Italian pathological speech in

real time through algorithms integrated in non-invasive tools and digital applications. In particular, it appears essential to explore the acoustic cues of oesophageal voices in order to enrich the spectrogram and the phonetic characteristics of spontaneous speech to achieve an adequate level of intelligibility and auditory pleasantness.

The organisation of the paper is as follows: the state of the art focuses on alaryngeal speech and on its phonetic cues as analysed in previous works. Following these theoretical premises, we present our data relating to the sociolinguistic profile and clinical history of our target speaker along with the research protocol regarding data collection, annotation and the methodology used to extract the main acoustic features from a sub-sample of our speaker's production. The results present a qualitative analysis of the main acoustic features of a total laryngectomized speaker, with a particular emphasis on vowels' formants, fundamental frequency and voice quality (jitter, shimmer, HNR). These results are discussed with respect to previous literature on oesophageal speech. The different applications of these results in the improvement of the latest generation algorithms to enhance the speech quality of patients suffering from a total laryngectomy are listed below. In this sense, an overview of the current state of the art for speech enhancement and reconstruction technologies is also provided and discussed in order to open the path for the creation of further tools for improving the voice quality (and with that the quality of life) of laryngectomized patients. In conclusion, the future perspectives of this research can help in the development of specific technologies for laryngectomized speakers, thus underlining the need for an interdisciplinary collaboration for creating new generation algorithms.

2. Laryngectomy and Speech: A challenge for Clinical Phonetics

Total or partial laryngectomy is usually a consequence of neck cancer, whose genesis is supposed to be multifactorial (Brouha et al., 2005). Casper and Colton (1998) have underlined that laryngeal cancer is mostly diagnosed in 60-years old heavy smoker men with moderate alcohol intakes (but also cf. the contrastive findings in Goldstein & Irish, 2005). Neck cancers are usually formed by squamous cell carcinomas who need to be removed surgically, with a following localized chemotherapy or a combined chemo-radiotherapy. However, chemo- and radiotherapy also has side-effects with respect to the patients' mucous (Brosky, 2007), subsequently leading to negative influences on their speech therapy and rehabilitation.

After a total laryngectomy, it is possible for patients to speak again by choosing one of the three main ways (Štajner-Katušić et al., 2004): external speech aids (e.g., laryngophones), voice prosthesis or trachea-oesophageal speech (TES) and oesophageal speech (ES). While the first two ways rely upon technological enhancements, oesophageal speech requires regular practice even for proficient speakers, who need to learn how to fill the oesophagus with air and then expel it in a voluntary and controlled fashion. Bressmann (2010) describes this process, by emphasizing that speakers have two main ways to insufflating air into the upper oesophagus; either by using an inhalation manoeuvre or by injecting air using active pressure build-up in the oral cavity by manoeuvres such as air swallows, glossopharyngeal pumping, or forceful articulation of an unreleased plosive such as [k]' (Bressmann, 2010, p. 510). Graham (2005) points out that fluent speakers may use a combination of both these ways and could reach a production up to ten syllables upon a single insufflation. These insufflations cause a noise in the spectrogram above 5,000 Hz, but may also cause other forms of disturbance throughout the recording.

Studies on alaryngeal speech have taken into account the comparison between the acoustic properties of voice and sounds as produced by speakers with partial or total laryngectomy. A huge part of the research has been devoted to addressing the ways of producing voiced consonants in absence of vocal folds vibrations (Christensen et al., 1978, and also Štajner-Katušić et al., 2004 for a review). However, this part is not addressed here as it is out of the scope of this study.

In contrary to voice produced with external devices such as electro-larynx or laryngophone, both TES and ES produce the voice in the pharyngo-oesophageal (PE) segment. Bressmann (2010, p. 510) describes the voice of ES speakers as rough and with a low pitch, due to the articulatory dynamics involved during speaking. The air stream is originated into the oesophagus, with the PE segment

located between the third and sixth cervical vertebra (Štajner-Katušić et al., 2004, p. 195). Thus, the speaker needs to insufflate the upper oesophagus that is present below the level of the upper oesophageal sphincter and then expel the air in a controlled fashion (Bressman, 2010, p. 510). As a result, the speakers have little control over the volume and the pitch. Furthermore, from an acoustic phonetic perspective, these dynamics of speech production change the length and shapes of the filter and, in the traditional source-filter model (Fant, 1960), this results in a modification of the acoustic cues traditionally associated with speech.

Van Sluis et al. (2018, p. 13) present a useful review of the main phonetic features associated with the comparison of ES and TES speakers. The authors include a primary outcome as fundamental frequency (F0), harmonics-to-noise ratio (HNR), and the percentage of voiced-ness (%voiced), whereas secondary acoustic parameters were jitter, shimmer, intensity, spectral tilt and maximum phonation type (MPT). Debruyne et al. (1994) have already discussed the role of F0 and HNR in differentiating TES and ES speakers from control groups (i.e., laryngeal speakers). The authors demonstrate that alaryngeal speakers show a lower pitch, and 'a flatter spectrum and a relatively higher level of energy below 4000 Hz' (Debruyne et al., 1994, p. 327). They also show the role of jitter and shimmer in differentiating among TES, ES and control speakers. This data is in line with Robbins's (1984) findings, who reported that jitter was 18.2% higher for ES speakers than in normal voices. As for spectral tilt, however, albeit its importance for distinguishing phonation types (e.g., Jackson et al., 1985) and stress or intonation prominence (Campbell & Beckman, 1997), its relevance for the study of alaryngeal speech has not been fully demonstrated, and other measurements have been preferred (see Kobayashi et al., 1995).

3. Material

For the purpose of performing a preliminary analysis of ES in Italian, we recorded a first target speaker we thereby referred to as ESO01, due to privacy reasons. ESO01 has volunteered in taking part in a recording session performed in an informal albeit sound-controlled setting in July 2020. Despite the Covid-19 pandemic, it was possible to perform the recording in a safe environment by respecting the security distances among participants so that ESO01 could speak without a face mask. The speaker was aware of being recorded and of the general purpose of the study. He signed an informed consent for the use of his recordings in an anonymized form. The consent form was designed in accordance to the ethical considerations for 'vulnerable populations' with the current Italian and Swiss legislation regarding audio recordings for clinical research (see also Powell, 2013, pp. 14-15).

A conversation lasting almost 1 hour was recorded with a TASCAM DR20 (sampling rate 44.1 KHz, 16 bit). The first two authors invited ESO01 to talk about his experience and how he reached his current oesophageal speech level as it was considered by doctors as an excellent goal for intelligibility and durability over time. ESO01 spontaneously introduced other topics in the conversation including episodes of his private life, which were not transcribed for privacy reasons. At the end of the informal recordings, ESO01 was also asked to read aloud three times a list of sentences with Italian diphthongs and cardinal vowels (Draetta, 2019). The list consisted in 26 different disyllabic target words, balanced for target vowel or diphthong, with the same metrical and intonational (i.e., adfirmative) structure: for instance, *Questa sarta non è brava* "This seamstress is not skilled" or *La paura è uscire dall'euro* "The fear is to get off euro".

For the present study, we used the 3' of read speech and 20' of spontaneous speech selected from the first part of the interview and was completely anonymized. The transcription and annotation was performed on two tiers of PRAAT. The first tier is devoted to the orthographic transcription by isolating the different portion of speech between silences. On the second tier, we annotated vowels by basing them on F2 transition (Di Paolo et al., 2011); only clearly recognisable vowels were annotated and phonologically transcribed.

We also isolated two different acoustic phenomena which are typical of ES on both the first and second tiers and labelled them as VOC (vocoid) and RIL (release) and shown in Figure 1.

The so-called vocoid has been identified as a short phonation (around 85 msec.) characterized by the presence of formants in a vowel-like fashion but without an identifiable vowel quality. This vocoid

has been found almost at the beginning of each long phonatory emission in particular, before complex sentence structures. In a similar fashion, the release has been found directly before or immediately after the vocoid and it is represented by an airstream emission similar to a frication noise, which could be interpreted as the inspiratory phase before speech.

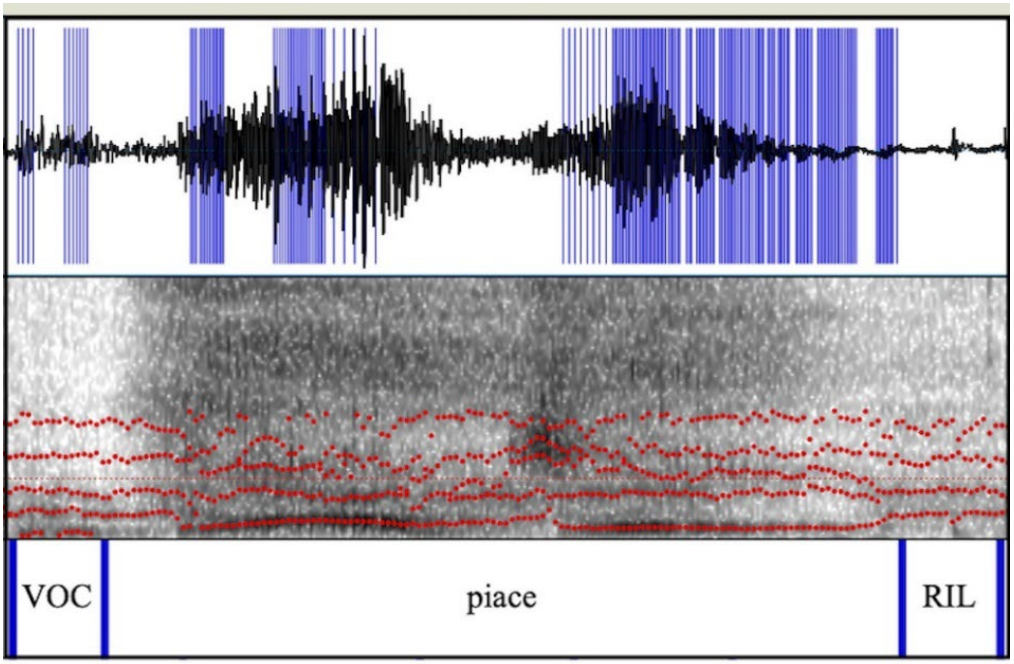

Figure 1. A spectrographic representation of the vocoid (on the left) and of the release (on the right) as realized during spontaneous speech.

4. Methodology

The audio corresponding to the spoken and spontaneous speech has been tagged on Praat, by isolating the major cues on two tiers called Words and Phones while vowel spaces have been realized with Visible Vowels (Heeringa & van de Velde, 2018). Vocoids and air release have been tagged as VOC and RIL, respectively, in the second tier together with the target vowels /a, e, ɛ, i, o, u/. No instances of /ɔ/ have been found, which could be justified by the regional (north-western) Italian accent of our target speaker.

Vocoids (VOC) occur at the beginning of each sentence, and they have been tagged by examining the starting and ending point of the formants-like transition. The vocoid almost always lasts until the beginning of the sentence. It rarely happens that the formantic components finish before the sentence leaving a pause of about 40 ms. Vocoids were annotated for both read and spontaneous speech.

Being visually temporally stable in terms of formantic components (but see also the analysis below), the extraction of the first three formants' values (plus the intensity) on the vocoids and on the vowels was done by adopting the following setting: time step 10 ms, Gaussian-like analysis window of 50 ms, maximum formant value, 5500 Hz, pre-emphasis 50 Hz and LPC coefficients using the algorithm by Burg (cf. Childers, 1978 and Press et al., 1992). Analysis was carried out on 410 samples for read speech and 5480 for spontaneous speech.

The analysis of air releases (RIL) was conducted only for spontaneous speech because they are more indicative in the context of continuous speech, where ESO01 adopts personal strategies to manage the emission of multiple sentences in sequence. In continuous speech, the RIL is almost always followed by another VOC and they both precede the whole sentence. RILs have been identified by a tag always beginning with the conclusion of the sentence and lasting until harmonic components are present. The air release visually presents a constant formants value over time. As a result, we proceed with the same methodology that was used for vocoids thus, extracting formants from 8718 time-samples. The extraction of the first three formants and of fundamental frequency was done by the second tier Phones as average values and envelope

with 7 points, by means of a specific Praat script[1]. Consequently, the extrapolation of the formants' values from 1207 vowels in spontaneous speech and 99 in read one with a total amount of 1306 tokens was done. Voice quality features have been extracted on 119 /a/ vowels produced during the spontaneous conversation. We used a specific Praat script[2] with cross-correlation method and the following setting; time step 10 ms, minimum pitch 50 Hz, silence threshold 10 ms, number of periods per window, 4.5.

5. Analysis

5.1. Vocoids and air release

A first qualitative inspection of the data shows how the vocoids were equally present in both spontaneous and in read speech. However, a difference is observed between the two speaking styles considering the formant characteristics of the vocoids. The formant values of F1, F2 and F3 were more stable in spontaneous speech while the instances of the vocoid found in read speech can be divided into two main groups (see Figures 2).

The mean values of the vocoids formant for spontaneous speech are 152.177 Hz (St. Dev. 16.423 Hz) for F0, 728.422 Hz (St. Dev. 93.819 Hz) for F1, 1707.81 Hz (St. Dev. 59.97 Hz) for F2, 2776.03 Hz (St. Dev. 109.05 Hz) for F3. The intensity as calculated at the midpoint of the interval is 56.765 dB (St. Dev. 1.547 dB). The major variability of the vocoid in read speech could be explained from a communicative point of view linked to the nature of the tasks, as the target speaker couldn't use his own words in the read speech. Therefore, he could not start the sentences with vowels /a/ or /e/, like he usually does in spontaneous speech. The lack of this possible articulatory strategy for starting the phonatory segments, could be responsible of the higher variability of the vocoid in read speech. As we have seen in Fig. 2, vocoid's formants in read speech are distributed between two main clusters, which are worth further investigations (e.g., in relation with the phone at the beginning of each sentence).

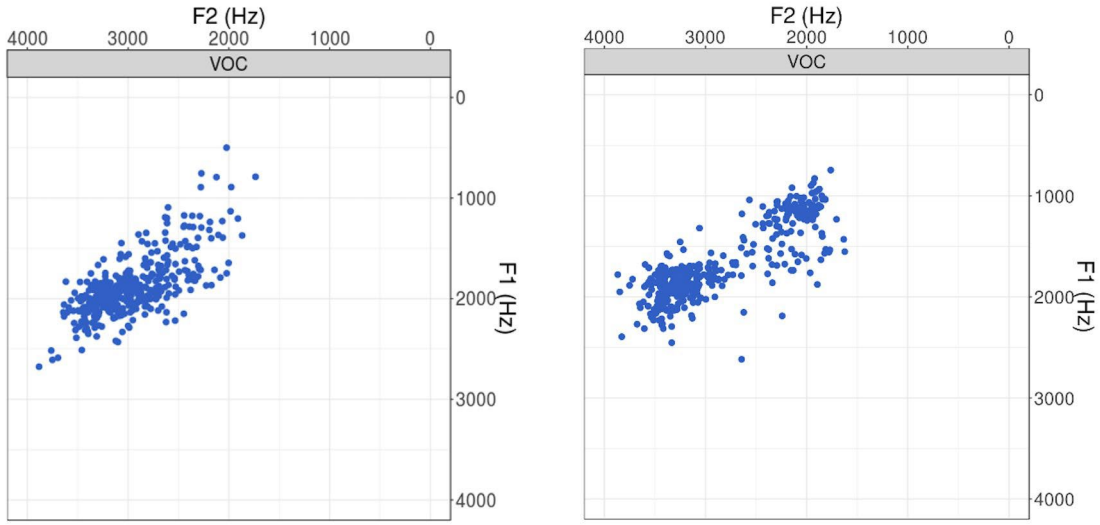

Figure 2. The concentration of formant values in a F1/F2 plot in read speech (right) and in spontaneous speech (left).

The boxplots in Figure 3 show the distribution of formants and intensity values through our whole corpus, that is by considering together spontaneous and read speech. It is evident that data for F1 and F3 present a higher degree of variability, whereas values for F2 and intensity (with the exception of one outlier) show a narrow distribution around the mean value.

[1] Praat script by Mietta Lennes (modified by Chiara Meluzzi for Italian data): http://www.helsinki.fi/~lennes/praat-scripts/.

[2] Public script by David R. Feinberg (https://osf.io/dbrpf/).

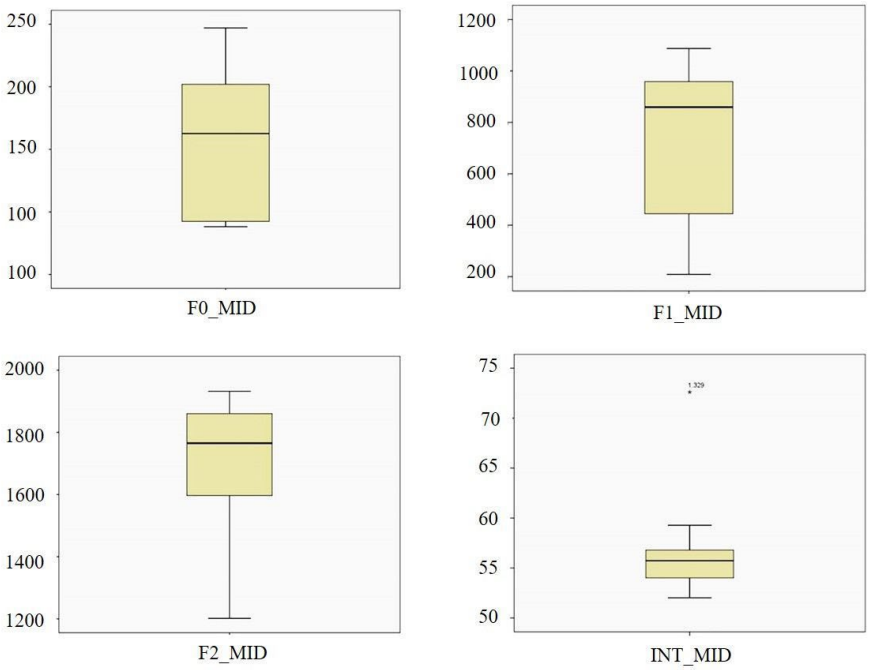

Figure 3. Boxplots of the distribution of formants and intensity values in the vocoids as calculated at the midpoint on the whole dataset.

RIL shows a great stability in its formants' values with similar distribution in both spontaneous and read speech (see Fig. 4). The mean values of RIL for spontaneous speech are of 341.954 Hz (St. Dev. 4.93) for F1, 1983.82 Hz (St. Dev. 2.27) for F2 and 2966.504 Hz (St. Dev. 4.125) for F3.

The fact that both the vocoid and the release show stable behaviour could spark further interest in the improvement of algorithms for the identification and development of tools specifically dedicated to oesophageal speakers (see below for a further discussion).

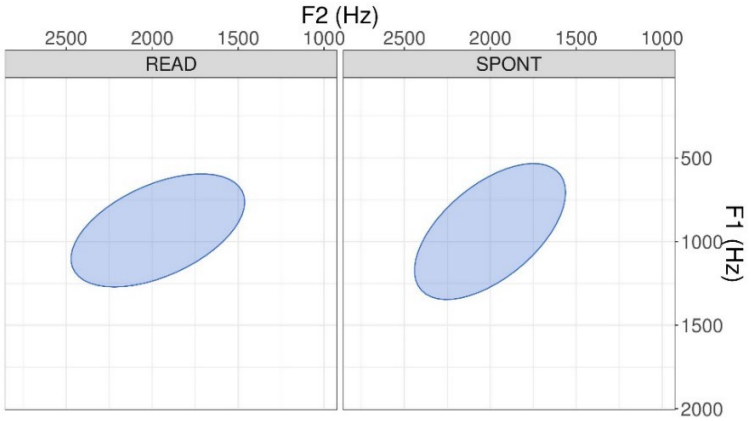

Figure 4. The formants values of RIL in a F1/F2 plan for read speech (on the left) and spontaneous speech (on the right).

5.2. Vocoids and air release in spontaneous speech

We focused only on vowels produced during spontaneous speech, in order to further analyse their variability. We considered vowels' trajectories through different time-points together with the values of each vowel's formants taken at the midpoint, in order to combine a static and dynamic vowel analysis. We had to exclude vowel /u/ from the present analysis since F1 and F2 values were not clearly distinguished automatically, and they appeared quite overlapping between each other, in a way that has already been noticed in other cases of Italian pathological speech (cf. Meluzzi, 2021, p. 423).

Table 1 shows the overall mean results of vowels' formants as extracted at the midpoint. It is possible to notice that these values are higher with respect to a laryngeal male speaker (see also the discussion) but with a certain consistency in particular for what it concerns the relationship between F1 and F2 across vowel types.

	F0	F1	F2	F3
/a/	136.55 Hz (Std. 5.416 Hz)	679.53 Hz (Std. 5.637 Hz)	1506.31 Hz (Std. 12.47 Hz)	2701.52 Hz (Std. 25.1)
/e/	184.49 Hz (Std. 7.19 Hz)	524.12 Hz (Std. 4.84 Hz)	1921.81 Hz (Std. 20.03 Hz)	2677.14 Hz (Std. 25.9 Hz)
/ɛ/	172.36 Hz (Std. 15.32 Hz)	526.447 Hz (Std. 9.296 Hz)	1982.57 Hz (Std. 57.93 Hz)	2686.88 Hz (Std. 74.161 Hz)
/i/	141,907 Hz (Std. 7.47 Hz)	371.04 Hz (Std. 8.92 Hz)	1761.498 Hz (Std. 41.11 Hz)	2754.73 Hz (Std. 43.34 Hz)
/o/	159.23 Hz (Std. 6.577 Hz)	545.78 Hz (Std. 4.56 Hz)	1293.327 Hz (Std. 26.097 Hz)	2733.189 Hz (Std. 27.77 Hz)

Table 1. Formants mean values and standard deviation for target vowels.

We also visually inspected the relationship between F2 and F3 as shown in Fig. 5. For this analysis, we excluded /ɛ/ instances since they were quite limited in our sample (50 tokens). From the graph, it is evident that the relative distribution of the two formants is quite linear with the partial exception of /i/ showing a greater variability. This justifies the claim that despite the variability, the relative difference between formants is maintained and it could represent a robust parameter for perceptually differentiating among vowels in alaryngeal speech. However, all vowels are distributed along the diagonal of the quadrant suggesting that despite the different variability, the difference between F3 and F2 remains almost constant. This could also contribute to the maintenance of intelligibility of the different vowels from a perceptual perspective.

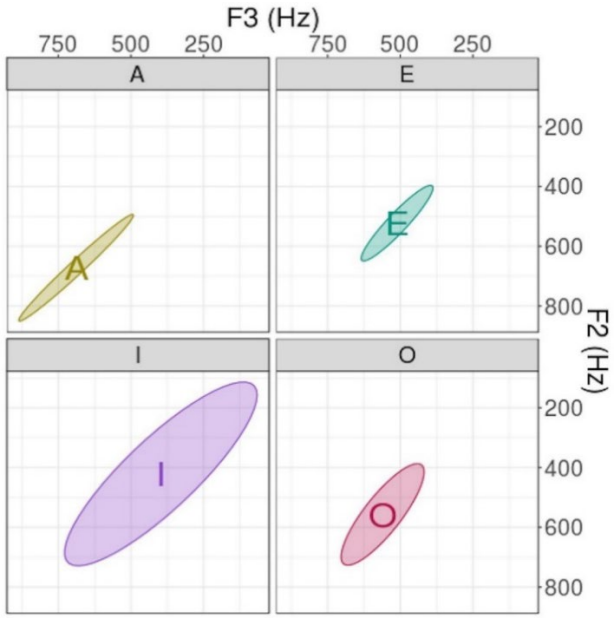

Figure 5. The relative variation of F2 and F3 in the four target vowels /a, e, i, o/.

The examination of the dynamic variation of fundamental frequency during the articulation of vowels /a/, /e/ and /o/ (Fig. 6) evidently predicted the stability of the curve between the second and fourth target point, whereas a major increase in the F0 contour has been observed in the two final time-points.

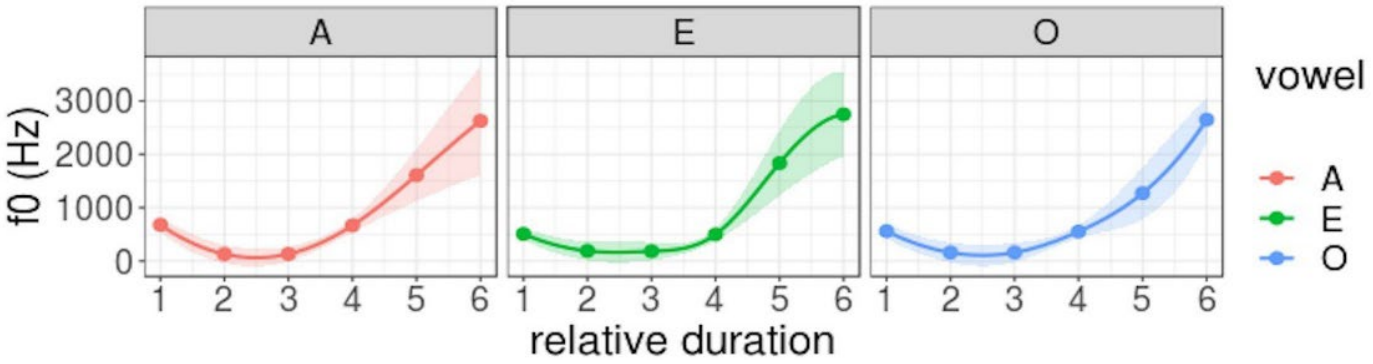

Figure 6. The dynamic variation in time of F0 in vowels /a/, /e/ and /o/ and the standard deviation thresholds.

The same pattern is observed for the target vowels /a, e, i, o/ in both F1 and F2 dynamic variation in spontaneous speech (see Fig. 7). Further analysis is required to take into account the phonological characteristics of the preceding and following consonants but from these preliminary dynamic patterns it appears that ESO01 manages to reach the acoustic target for each vowel and to maintain this target into a steady-state.

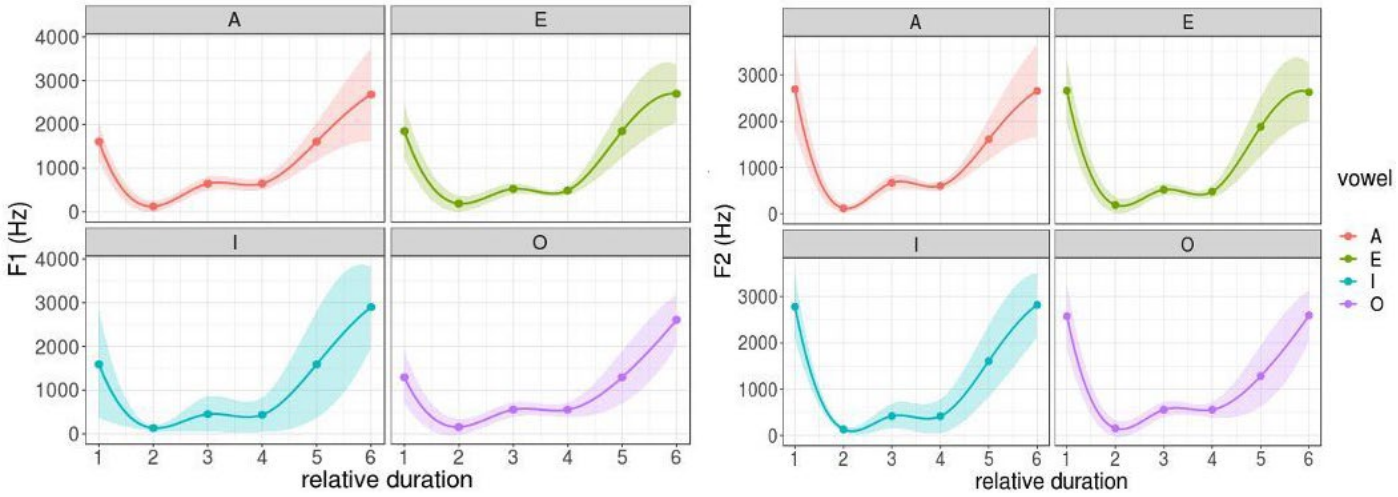

Figure 7. The dynamic variation of F1 (right) and F2 (left) in the target vowels in spontaneous speech.

5.3. Jitter, Shimmer and Harmonic to Noise Ratio (HNR)

As stated in the methodology, jitter, shimmer and NHR on all vowels /a/ produced in spontaneous speech is extracted with a sampling every 10 msec. The boxplots of the mean values of these parameters are shown in Fig. 8.

The boxplots highlight the presence of many outliers in particular for jitter and shimmer values. The mean value of jitter is 3.035 dB (St. Dev. 0.49 dB), whereas for shimmer, the value is 0.023 dB (St. Dev. 0.003 dB). It should be observed that the values of both jitter and, most of all, shimmer are far below the threshold for pathology when compared with established value of 0.350 dB according to the literature. Conversely, HNR values appear more uniformly distributed with a mean of 2.26 dB (St. Dev. 0.24 dB).

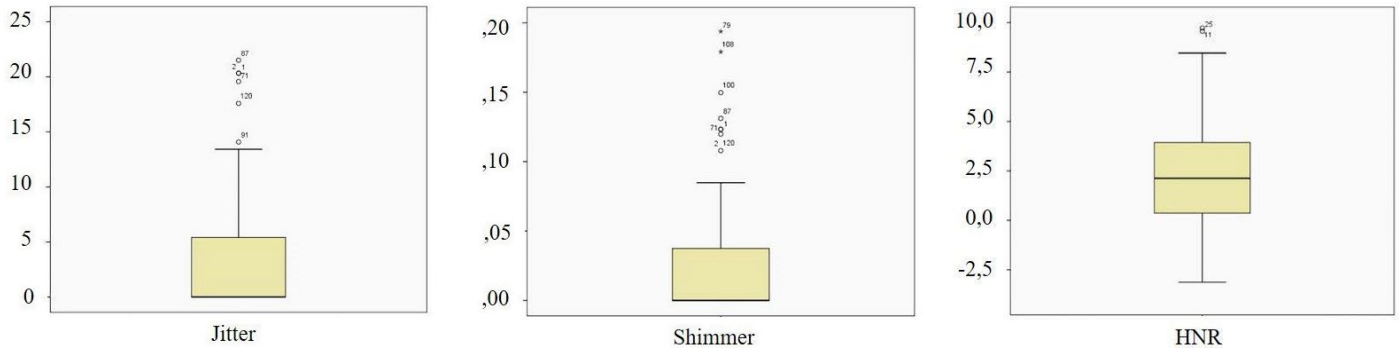

Figure 8. The boxplots of the distribution of jitter, shimmer and HNR values on vowel /a/.

6. Discussion of the results

6.1. Vowel formants

As stated at the beginning of this paper, this work represents the first preliminary investigation on alaryngeal Italian speech with the purpose of defining algorithms and technologies for speech recognition and enhancement in case of such pathological speech. In the following section, we will discuss the applications of our novel findings along with the possible linguistic implications of our results by making a parallel with previous findings attested in the literature.

Previously, it has been noted that the reduction in the length of the speech resonator (i.e., the laryngeal channel) leads to an increase of all vowel formants values in alaryngeal speech compared to average laryngeal one. The estimated variation was around 123 Hz for F1, and of 320 Hz for both F2 and F3 (Sisty & Weinberg, 1972, p. 443). However, recent works on different languages have demonstrated how this pattern of variation is not always set across all vowel types. The vowels /a/ and / ɛ/ are the most variables vowels, whereas a minor or null variation has been found for vowels /i/ and /y/ between alaryngeal and laryngeal speakers (cf. Esen Aydinli et al., 2019). Due to the lack of previous indication on vowels formants produced by an Italian oesophageal speaker in spontaneous speech productions, we make a comparison between the mean values of ESO01 and the formants values of an average male Italian speaker as reported in the literature (cf. Giannini & Pettorino, 1992); it should be noted that the speaker considered by Giannini & Pettorino (1992) was from central Italy even if it is claimed that he spoke without any clearly identifiable regional accent.

	ESO01		**Average Italian male speaker (Giannini & Pettorino, 1992)**	
	F1	**F2**	**F1**	**F2**
/a/	679.53 Hz	1506.31 Hz	750 Hz	1500 Hz
/e/	524.12 Hz	1921.81 Hz	350 Hz	2100 Hz
/ɛ/	526.447 Hz	1982.57 Hz	550 Hz	1750 Hz
/i/	371.04 Hz	1761.498 Hz	250 Hz	2250 Hz

Table. 2. Formants values for Italian vowels of ESO01 compared with the values reported in literature produced by a male speaker taken from the literature.

Table 2 demonstrates no difference between ESO01 and an average non-pathological male speech for vowel /a/. Conversely, vowels /e/ and /i/ show major differences, albeit not always in the sense of an increase in the mean values of formants as in case of our alaryngeal speaker. Indeed, it appears that ESO01 has higher formant values for both F1 and F2 only for vowel /o/, whereas a lower F2 value for /e/ and /i/. A lower formant value is also observed for the F1 of /ɛ/. It is evident that this similarity ought to be confirmed by a larger quantitative study, but it is worth trying to explain these preliminary qualitative results.

This variance could be differently explained. Firstly, it is evident that the lack of a control speaker of the

same sociolinguistic background constitutes a limitation to the present study. This means that comparing with the mean values reported in the literature is not sufficient, and certainly it would be necessary to provide a more precise parallel with a control speaker of the same age and geographical origin to avoid a possible influence of the dialect of our target speaker in his spontaneous speech. Taking it for granted, it is, however, clear that ESO01's formants are similar to a typical laryngeal speaker and that the differences are not always in an increase in the mean values. The picture that emerges is of a lower F2 and a higher F1 indicating a more posterior and higher recognition of the target vowels. This alludes to a reduction of the vowel space of our target speaker. Furthermore, the higher tongue position during phonation is confirmed by previous studies on languages other than Italian as reported in the state of the art. The posterior articulation, however, contrasts with previous studies (e.g., Cervera et al., 2001) who conversely reported a more frontal articulation, especially for /i/.

6.2. Voice quality in oesophageal speakers: Technologies for speech enhancement

Although ESO01 speaks with sufficient intelligibility to lead an almost normal life, his oesophageal speech is extremely deteriorated when compared to that of a healthy speaker. ESO01 could use a throat microphone (laryngophone) to reduce physical effort, but by personal choice he prefers to speak without aids.

Throat microphones (Cohen et al., 1984; Liu & Ng, 2007, 2009; Sahidullah, 2017) were successfully applied for voice activity detection and speaker recognition (Sahidullah et al., 2016-2017), and if combined with acoustic microphones they have been used to partially reconstruct the spectrum of the speech (Zheng et al., 2003; Erzin, 2009; Shahina & Yegnanarayana, 2007; Turan, 2018) in controlled conditions. Moreover, it is psychologically important for the users to have an acoustically acceptable voice. Nonetheless, the solutions actually available in the market are still far from being good from both a qualitative and a comfort point of view. Consequently, it is common for the speakers to choose the oesophageal speech technique. Hardware products are still limited to old-styled tools like artificial larynx and simple signal amplifiers that do not take into account refined features related to intelligibility and timbre (e.g. Luminaud, Griffin Laboratories, and Atos Medical, or UltraVoice). On account of the development of machine learning techniques, the research in the field of enhancement and reconstruction of degraded speech has increased enormously in recent years. Improving the oesophageal speech is possible, but algorithms must be robust and fast enough to allow implementing the voice signal in real-time (or with a minimal delay) to be incorporated into aid technologies for daily interaction. Nevertheless, none of the newly developed algorithms appear to be yet implemented in products which matched real-time usability and non-invasive external aids. Moreover, it can be seen from the previous state of the art that the research suffers from a lack of in-depth phonetic studies and specific corpora of data regarding speakers with permanent damage or removal of vocal cords.

Recent advancements in artificial intelligence successfully increased the intelligibility of NAM (Non-Audible Murmur) microphones (Nakajima et al., 2006) by means of Generative Adversarial Neural Networks applied to NAM-To-Whisper and Whisper-To-Speech tasks (Shah & Patil, 2020; Zhou et al., 2012; Pascual et al., 2019). Other techniques make use of Hidden Markov Models and Gaussian Mixture Models to perform speaker recognition (Patel et al., 2019) and speech enchantment. Nevertheless, none of these algorithms has yet been implemented in products for real-time usage. Indeed, the problem of Generative Adversarial Networks (GAN) is that they are significantly difficult to train in terms of computational complexity (Pascual et al., 2019). In GAN architectures, a generative adversarial neural network is trained with power spectrum and other high-level spectral features to produce a speech signal from a NAM signal by finding the relation of the respective feature vectors (Turan, 2018). It must be emphasized that all these high-level features are used in many different signal processing fields and often do not take into account the specific peculiarities of degraded spontaneous speech going down to the phonetic level. In the case of Deep Neural Network, the procedure is almost the same, but it requires the network to have a higher number of hidden layers and, thus, allows to learn more complex relations between source and target spectral feature vectors with a consequent higher

computational cost. Gaussian Mixture Models were successfully applied in tasks as Voice Conversion (Stylianou, 1996) leading to successful results also in speech enhancement tasks if combined with Maximum Likelihood Estimation (MLE) as stated in (Toda & Shikano, 2005). In GMM architectures for Whisper-To-Speech tasks, pairs of NAM and speech signals are fed to two GMM for spectral estimation and pitch estimation. In this case, the features and the power spectrum of both signals are previously extracted, and the model tries to find the mapping between the source and target feature vectors. Finally, the Hidden Markov Model (HMM) research approach is divided into more steps than the previous techniques. First, source and target parameters are modelled by context-dependent phone-sized HMM. Then, an HMM recognition is applied on the input feature vectors, and a third HMM is used for the synthesis of the speech (Tran et al., 2009), but they typically require more computational time. The quasi-real-time of oesophageal Italian speech reconstruction could exploit more than all a modified CELP codec approach designed for whispered speech as stated in (Sharifzadeh et al., 2010) instead of replicating Generative Adversarial Networks (GAN) and traditional Neural Networks (NNs). CELP is a linear predictive model of speech production used in a short-term predictor to model the spectral envelope of the speech. ESO01's voice has many similarities with the whispered speech. Vowels present clear formant values, as needed for pitch reconstruction and estimation, and the presence of air emissions is due to the phonation modality, which differs from laryngeal healthy speech.

7. Conclusions and future works

Total laryngectomy severely impairs communication, even if personal motivation could lead to a comprehensible speech. In this study, we have presented a preliminary analysis of one Italian oesophageal speaker (ESO01) and contrary to previous works, we have focused not only on sustained vowels or controlled speech, but also on vowels as produced during a long spontaneous conversation in an informal setting. Albeit the obvious sample limitation of this work, it was possible to provide some first considerations on the acoustic characteristics of ES in real communication.

A general intelligibility of vowel quality, as resulting from a relatively homogenous distribution of formants values and on the maintenance of distinction across vowels, is counterbalanced by a voice quality characterized by the noise implicit in the articulatory mechanism of oesophageal speech. The results allow us to hypothesize new approaches for the development of speech enhancement algorithms. First of all, the constant and characteristic behaviour of vocoids and air releases suggest the possibility to automatically remove these specific noises in order to improve intelligibility, for example, making the beginning of the sentence clearer.

Secondly, oesophageal vowels are characterized by irregular behaviour with respect to healthy voices due to the lack of vocal cords, it seem essential to rethink speech enhancement aids based on segmental and phonetic features, pitch estimation, and phonetic reconstruction, as opposed to general amplification and de-noising techniques. A less generalized approach can be customized on the specific oesophageal speech of the patient, for example, by means of the cited NAM microphone and speech enhancement computational models. Expanding the database will be then useful also to highlight general constant behaviours in Italian oesophageal speech, plan new targeted strategies for automatic speech enhancement, and to validate them from the point of view of comprehensibility and voice pleasantness. Furthermore, findings on alaryngeal speech should be integrated within those models aiming at creating and validating indexes for the measure and evaluation of pathological voices, like the Acoustic Voice Quality Index (AVQI, cfr. Maryn et al., 2020, and, for Italian, Fantini et al., 2021).

Finally, these results could be helpful in clinical practice. In particular, the analysis of vocoids and release could help speech therapists in coordinating insufflating movements and speech in laryngectomized speakers. Moreover, a visible output of their speech could also be helpful for speakers during their rehabilitation, as it has been largely demonstrated by clinical phonetic experiments conducted on children (e.g., Preston et al., 2016).

Further research should confirm these findings with a wider sample, but also through perceptual tests. Other segmental and supra-segmental features should also be considered, by analysing temporal features and also vowel variability in different phonological contexts. For instance, it could be interesting to address the issue of F2/F3 variability in case of nasals or nasalized vowels, as well as voicing qualities in consonants from both an acoustic and a perceptual perspective. Finally, also concerning speaker's intelligibility and its measurements through the AVQI test, it could also be interest to record our target speaker with the facemask, in order to verify whether and to what extent the comprehension of his speech is negatively affected, as it has been shown by the recent study conducted by Ribeiro et al. (2020), among many others.

Acknowledgements

The paper has been jointly written by the four authors. However, for the requirements of the Italian Academy, Chiara Meluzzi is responsible for sections 1, 2, 3, 4.1 and 5.1, other than data collection and their phonetic annotation. Sonia Cenceschi is responsible for sections 4.2, 5.2 (together with Francesco Roberto Dani) and 6, other than data collection and the automatic extraction of acoustic parameters. Finally, Alessandro Trivilini is responsible for the revision of the work, and for the overall project.

References

Bressmann, T. (2010). Speech disorders related to head and neck cancer: Laryngectomy, glossectomy, and velopharyngeal and maxillofacial deficits. In J. S. Damico, N. Müller, & M. Ball (Eds.), *The handbook of language and speech disorders* (pp. 497-526). Wiley-Blackwell.

Brosky, M. E. (2007). The role of saliva in oral health: Strategies for prevention and management of xerostomia, *The Journal of Supportive Oncology, 5*(5), 215-225.

Brouha, X., Tromp, D., Hordijk, G. J., Winnubst, J., & De Leeuw, R. (2005). Role of alcohol and smoking in diagnostic delay of head and neck cancer patients. *Acta Oto-Laryngologica, 125*(5), 552-556.

Campbell, N. & Beckman, M. (1997). Stress, prominence, and spectral tilt. In A. Botinis, G. Kouroupetroglou, & G. Crayiannis (Eds.), *Intonation: theory, models and applications* (pp. 67-70). European Speech Communication Association.

Casper, J. K., & Colton, R. H. (1998). *Clinical manual for laryngectomy and head & neck cancer rehabilitation*. Singular.

Cervera, T., Miralles, J. L., & González-Alvarez, J. (2001). Acoustical analysis of Spanish vowels produced by laryngectomized subjects. *Journal of Speech, Language and Hearing Resarch, 44*(5), 988-96.

Childers, D. G. (ed.). (1978). *Modern spectrum analysis*. IEEE Computer Society Press.

Christensen, J. M., Weinberg, B., & Alfonso, P. J. (1978). Productive voice onset time characteristics of esophageal speech. Journal of Speech. *Language and Hearing Resarch, 21*(1), 56-62.

Cohen, A., Van Den Broeckero, M. P., & Van Geel, R. C. (1984). A study of pitch phenomena and applications in electrolarynx speech. *Speech and Language, 11*, 197-248.

Cummings, L. C., & Cooper, G. S. (2008). Descriptive epidemiology of esophageal carcinoma in the Ohio Cancer Registry, *Cancer detection and prevention, 32*(1), 87-92.

Esen Aydinli, F., Kulak Kayikci, M. E., & Suslu, N. (2019). Temporal and Frequency Characteristics of Turkish Vowels in Laryngectomized Speakers: Preliminary Study. *Medeniyet Medical Journal, 34*(2), 149-159.

Debruyne, F., Delaere, P., Wouters, J., & Uwents, P. (1994). Acoustic analysis of tracheoesophageal speech. *The Journal of Laryngology and Otology*, *108*, 325-328.

Doyle, P. C., & Finchem, E. A. (2019). Teaching esophageal speech: A process of collaborative instruction. In P. C. Doyle (Ed.), *Clinical Care and Rehabilitation in Head and Neck Cancer* (pp. 145-161). Springer.

Di Paolo, M., Yaeger-Dror, M., & Wassink, A. B. (2011). Analyzing vowels. In M. Di Paolo, & M. Yaeger-Dror (Eds.), *Sociophonetics. A student's guide* (pp. 87-106). Routlege.

Draetta, L. (2019). *Dittonghi e iati nella pronuncia di bambini biellesi: un'analisi sociofonetica* [MA thesis]. Università di Pavia.

Erzin, E. (2009). Improving throat microphone speech recognition by joint analysis of throat and acoustic microphone recordings. *IEEE transactions on audio, speech, and language processing, 17*(7), 1316-1324.

Fant, G. (1960). *The acoustics of speech*. Mouton De Gruyter.

Fantini, M., Maccarini, A. R., Firino, A., Gallia, M., Carlino, V., Gorris, C., Spadola Bisetti, M., Crosetti, E., & Succo, G. (2021). Validation of the Acoustic Voice Quality Index (AVQI) Version 03.01 in Italian. *Journal of Voice*, S0892-1997(21)00092-8 [Advance online publication].

Giannini, A., & Pettorino, M. (1992). *La fonetica sperimentale*. Edizioni Scientifiche Italiane.

Goldstein, D. P., & Irish, J. C. (2005). Head and neck squamous cell carcinoma in the young patient. *Current Opinion in Otolaryngology and Head and Neck Surgery, 13*(4), 207-11.

Graham, M. S. (2005). Taking it to the limits: Achieving proficient esophageal speech. In P. C. Doyle, & R. L. Keith (Eds.), *Contemporary considerations in the treatment and rehabilitation of head and neck cancer* (pp. 379-430). Pro-Ed.

Heeringa, W., & Van de Velde, H. (2018). Visible Vowels: A Tool for the Visualization of Vowel Variation. In I. Skadiņa, & M. Eskevich (Eds.), *Proceedings of CLARIN Annual Conference 2018, Pisa, Italy* (pp. 124-127). CLARIN.

Jackson, M., Ladefoged, P., Huffman, M., & Antoñanzas-Barroso, N. (1985). Measures of spectral tilt. *The Journal of the Acoustical Society of America*, *77*, S86 [2:49, MM8].

Kobayashi, N., Horiguchi, S., Baer, T. (1985) Aerodynamic and acoustic characteristics of the voicing distinction in electronic larynx speech. *The Journal of the Acoustical Society of America*, *77*, S86 [3:13, MM10].

Liu, H., & Ng, M. L. (2007). Electrolarynx in voice rehabilitation. *Auris Nasus Larynx, 34*(3), 327-332.

Liu, H., Ng, M. L. (2009). Formant characteristics of vowels produced by Mandarin esophageal speakers. *Journal of Voice, 23*(2), 255-60.

Maryn Y, Corthals P, Van Cauwenberge P, et al. (2010). Toward improved ecological validity in the acoustic measurement of overall voice quality: Combining continuous speech and sustained vowels. *Journal of Voice, 24*(5), 540-555.

Meluzzi, C. (2021). Sound Spectrography. In M. Ball (Ed.), *Handbook of Clinical Phonetics* (pp. 418-443). Routledge.

Nakajima, Y., Kashioka, H., Campbell, N., & Shikano, K. (2006). Non-audible murmur (NAM) recognition. *IEICE Transactions on Information and Systems, 89*(1), 1-4.

Pascual, S., Serrà, J., & Bonafonte, A. (2019). Towards generalized speech enhancement with generative adversarial networks. arXiv preprint. In G. Kubin, & Z. KačičProc (Eds.), *Proceedings of Interspeech 2019, Graz, Austria* (pp. 1791-1795). International Speech Communication Association.

Patel, M., Parmar, M., Doshi, S., Shah, N., & Patil, H. A. (2019). Novel Inception-GAN for Whisper-to-Normal Speech Conversion. In M. Pucher (Ed.), *Proceedings of 10th ISCA Speech Synthesis Workshop (SSW 10), Vienna, Austria* (pp. 87-92). International Speech Communication Association.

Powell, T. W. (2013). Research Ethics. In N. Muller & M. J. Ball (Eds.), *Research Methods in Clinical Linguistics and Phonetics. A practical guide* (pp. 10-27). Wiley-Blackwell.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge University Press.

Preston, J. L., Maas, E., Whittle, J., Leece, M. C., & McCabe, P. (2016). Limited acquisition and generalisation of rhotics with ultrasound visual feedback in childhood apraxia. *Clinical Linguistics & Phonetics, 30*(3-5), 363-381.

Ribeiro, V. V., Dassie-Leite, A. P., Pereira, E. C., Nunes Santos, A. D., Martins, P., & Irineu, R. de A. (2020). Effect of wearing a face mask on vocal self-perception during a pandemic. *Journal of Voice*, S0892-1997(20)30356-8 [Advance online publication].

Robbins, J. (1984). Acoustic differentiation of laryngeal, esophageal, and tracheo-oesophageal speech. *Journal of Speech and Hearing Research, 27*(4), 577-585.

Sahidullah, M., Gonzalez Hautamäki, R., Lehmann, Thomsen., D. A., Kinnunen, T., Tan, Z.-H., Hautamäki, V., Parts, R., & Pitkänen, M. (2016). Robust speaker recognition with combined use of acoustic and throat microphone speech. In N. Morgan (Ed.), *Proceedings of Interspeech 2016, San Francisco, USA* (pp. 1720-1724). ISCA.

Sahidullah, M., Thomsen, D. A. L., Hautamäki, R. G., Kinnunen, T., Tan, Z. H., Parts, R., & Pitkänen, M. (2017). Robust voice liveness detection and speaker verification using throat microphones. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26*(1), 44-56.

Shah, N. J., & Patil, H. A. (2020). Non-audible murmur to audible speech conversion. In H. A. Patil, & A. Neustein (Eds.), *Voice Technologies for Speech Reconstruction and Enhancement* (pp. 125-150). De Gruyter.

Shahina, A., & Yegnanarayana, B. (2007). Mapping speech spectra from throat microphone to close-speaking microphone: A neural network approach. *EURASIP Journal on Advances in Signal Processing, 087219*.

Sharifzadeh, H. R., McLoughlin, I. V., & Ahmadi, F. (2010). Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec. *IEEE Transactions on Biomedical Engineering, 57*(10), 2448-2458.

Sisty, N. L., & Weinberg, B. (1972). Formant frequency characteristics of esophageal speech. *Journal of Speech and Hearing Research, 15*(2), 439-448.

Štajner-Katušić, S., Horga, D., Mušura, M., & Globlek, D. (2004). Voice and Speech after Laryngectomy. *Clinical Linguistics & Phonetics, 20*(2/3), 195-203.

Stylianou, Y. (1996). *Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification* [PhD thesis]. Ecole Nationale Superieure des Telecommunications.

Toda, T., & Shikano, K. (2005). NAM-to-speech conversion with Gaussian mixture models. In I. Trancoso (Ed.), *Proceedings of Interspeech 2005, Lisbon, Portugal* (pp. 1957-1960). ISCA.

Tran, V. A., Bailly, G., Loevenbruck, H., & Toda, T. (2009). Multimodal HMM-based NAM-to-speech conversion. In. R. Moore (Ed.), *Proceedings of Interspeech 2009, Brighton, United Kingdom* (pp. 656-659). ISCA.

Turan, M. A. T. (2018). Enhancement of Throat Microphone Recordings Using Gaussian Mixture Model Probabilistic Estimator. *arXiv preprint*, arXiv:1804.05937.

van Sluis, K. E., van der Molen, L., van Son, R. J., Hilgers, F. J., Bhairosing, P. A., & van den Brekel, M. W. (2018). Objective and subjective voice outcomes after total laryngectomy: A systematic review. *European Archives of Oto-Rhino-Laryngology, 275*(1), 11-26.

Williams, S. E., & Watson, J. B. (1987). Speaking proficiency variations according to method of alaryngeal voicing. *The Laryngoscope, 97*(6), 737-739.

Zheng, Y., Liu, Z., Zhang, Z., Sinclair, M., Droppo, J., Deng, L., & Huang, X. (2003). Air-and bone-conductive integrated microphones for robust speech detection and enhancement. In J. Bilmes, & W. Byrne (Eds.), *IEEE Workshop on Automatic Speech Recognition and Understanding [St. Thomas, VI, USA]* (pp. 249-254). IEEE.

Zhou, J., Liang, R., Zhao, L., & Zou, C. (2012). Whisper intelligibility enhancement using a supervised learning approach. *Circuits, Systems, and Signal Processing, 31*(6), 2061-2074.