INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Long-Term Stability of Tracheoesophageal Voices

Klaske E. van Sluis[1,2], *Michiel W.M. van den Brekel*[1,2], *Frans J.M. Hilgers*[1,2], *Rob J.J.H. van Son*[1,2]

[1] Netherlands Cancer Institute, Amsterdam, the Netherlands
[2] ACLC, University of Amsterdam, the Netherlands
k.v.sluis@nki.nl

Abstract

Long-term voice outcomes of 13 tracheoesophageal speakers are assessed using speech samples that were recorded with at least 7 years in between. Intelligibility and voice quality are perceptually evaluated by 10 experienced speech and language pathologists. In addition, automatic speech evaluations are performed with tools from Ghent University. No significant group effect was found for changes in voice quality and intelligibility. The recordings showed a wide interspeaker variability. It is concluded that intelligibility and voice quality of tracheoesophageal voice is mostly stable over a period of 7 to 18 years.

Index Terms: pathological speech, tracheoesophageal speech, intelligibility, voice quality, long-term outcomes.

1. Introduction

Total laryngectomy (TL) refers to removal of the entire larynx as a treatment for advanced stage laryngeal cancer [2]. During the surgical procedure the airway and digestive tract are separated. With removal of the larynx the natural voice is lost. Voice rehabilitation is one of the most important goals after total laryngectomy [4]. In the early 1980s, insertion of a tracheoesophageal (TE) voice prosthesis (VP) was introduced [5]. A TE-VP is a one-way valve that is inserted in a puncture tract created between the trachea and esophagus. Airflow from the lungs to the mouth is thus reestablished. Henceforth, the patient is able to produce pulmonary driven speech again. The new voice source is the pharyngoesophageal segment (PES).

TE-speech is considered the gold standard in restoring communicative functioning after TL [2]. It is considered as the most natural way of voice restoration according to intelligibility, pitch, and range [6]. Success rates for acquiring TE-speech are reported up to 95% [7]. The reached endpoint in voice quality and speech intelligibility varies between patients [2]. Effective vibratory functioning of the PES is crucial in acquiring TE-speech. Knowledge about long-term voice outcomes of TE-speakers so far is scarce. There are some studies that include evaluations of TE-speakers on the long-term, up to 18 years post TL [8-12]. These papers, however, do not evaluate the groups of patients by follow-up time [8, 9, 11, 12]. Studies which consider long-term follow-up thus far only assess communication mode and quality of life [6].

In voice and speech assessment, a multidimensional approach is preferred. Acoustic, perceptual, aerodynamic, stroboscopic and self-assessment can be used to evaluate voice quality [13]. Substitute voices characteristically deviate from healthy speakers because of strong voice irregularities and require a well-thought-out approach [14]. As communication is mostly a perceptual matter, perceptual evaluations are considered the "gold standard". For substitute voices, judgments of experienced speech-language pathologists (SLP's) are considered as more consistent than judgments of naïve raters [11]. Various perceptual scales are applied in the literature to rate substitute voices. The IINFVo rating scale was specifically developed for substitute speech [15]. The five IINFVo scale parameters are: overall impression (I), impression of intelligibility (I), unintended additive noise (N), fluency (F) and voicing (Vo) [16]. Two of the rating scales, over-all impression and intelligibility, appear to be the most reliable [15-17] and are used in this study. The former refers to the acceptability or pleasantness of the voice (voice quality) and the latter refers to the clarity and understandability of words and sentences [16, 17]. During the last decade, automatic speech and voice analysis became feasible [18, 19]. Automatic analysis is promising in providing consistent ratings and for analyzing trends within a single speaker.

The present study aims to identify changes in TE-speech over time by analyzing perceptual and automatic evaluations of voice recordings.

2. Speech and methods

Speakers and speech recordings

The Netherlands Cancer Institute has a long history of speech collection for TL research. Recordings from 13 TL-patients, who participated in studies between 1996 and 2014, are included in the present study (all male, median age at treatment 55 years, range 44-75, all gave informed consent).

Inclusion was possible when voice recordings of the 149-word Dutch text with neutral content, "Tachtig dappere fietsers" [*Eighty brave cyclists*], were available from the same speaker with an interval of at least 7 years (T1 and T2, with 7-18 years in between) All speakers had undergone laryngectomy and used a Provox VP (Atos Medical. Hörby, Sweden). 27 recordings were made during the latter half of the 1990's (I), in 2007 (II), and in 2014 (III), in total 35 minutes of speech. No effort was made to ensure correct reading of the text so the actual words uttered vary somewhat. Speaker *UCX* did not complete the full text once and speaker *K9S* read a longer variant of the text once. For speaker *KRH*, there were recordings for all three periods (T1-T3).

Table 1. *Available patients/recordings, see text.*

Period		T1	T2
I	1996-1999	8	-
II	2007	5	7
III	2014	-	6

http://dx.doi.org/10.21437/Interspeech.2016-114

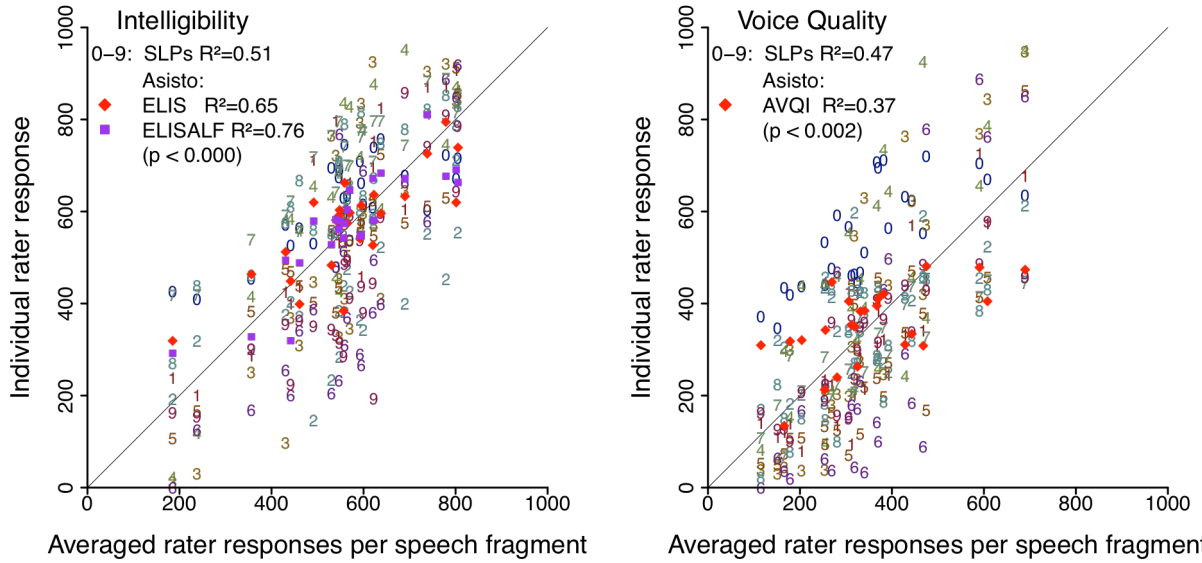

Figure 1: *Variation in perceptual evaluations and automatic scores in experiment 1. Numbers: individual expert raters; filled symbols: automatic ASISTO scores. Correlations are with average expert responses.*

The recordings were made as part of different studies, each using different equipment (see Table 2). For this study, recordings were digitized and converted to 44.1 kHz sampling rate and 16-bit Signed Integer PCM encoding (RIFF/WAVE). No audio compression had been used on the recordings.

Perceptual evaluation

Recordings were evaluated by ten experienced SLPs (experts), including one of the authors (KvS). Experts did the evaluations at home in a self-paced online listening experiment. At the time, experts were not informed about the details of the speakers. All experts were female, mean age 29.9 year (range 22-49). Eight were native speakers of Dutch. Two were native German speakers, who acquired Dutch as a second language. All experts were certified Dutch SLP's.

Evaluations were done using standard web browsers. There were two experiments. In experiment 1, the experts were asked to grade recordings of one single, long sentence as having better or worse speech intelligibility and voice quality. The experts used two slider rules as computerized visual-analog scales (VAS). In experiment 2, the same experts evaluated two pairs of short sentences from each speaker. The experts were asked to judge which version of the sentence in the pair was better and to what extent. The evaluation was again done using slider rules for speech intelligibility and voice quality. Experts could listen to the stimuli as often as they wanted. Stimuli were presented in pseudo-random order, different for each expert. The results were scored between 0-1000 (pseudo-continuous).

In experiment 1, a single, 16-word, sentence was used which resulted in 26 stimuli (13 for T1 and 13 for T2). In experiment 2, two different short sentences were used, 7 and 8 words long. Each pair in Experiment 2 was presented in both orders, T1/T2 and T2/T1, and both sentences were used. For each speaker there were four pairs, two sentences in two orders. For one speaker, *UCX*, one sentence of the recording was missing for T2. The missing sentence was replaced by another sentence. The results of this mixed pair are omitted here. In total there were 56 stimulus pairs in experiment 2, 2x13 sentence recording pairs in two orderings and 4 additional T2/T3 stimulus pairs for speaker *KRH*. Both experiment 1 and 2 were preceded with 5 practice items that were drawn from other speakers not in the test set.

Table 2. *Recording sessions.*

Period	Recorder	Microphone
1996-'99 [1]	Sony TCD-8†	AKG-c410
2007 [3]	Edirol Roland R1*	Sennheiser MD421
2014	Edirol Roland R09*	Samson Qv10e

†Digital Audio Tape (DAT) Deck. *Digital SD WAVE recorder

Automatic evaluation

The full 149 word recordings were automatically evaluated at the Department of Electronics and Information Systems, Ghent University with Automatic Speech analysis In Speech Therapy for Oncology (ASISTO) [20, 21]. Two applications for evaluating intelligibility were used, one using text aligned automatic speech recognition (ELIS), and one using alignment free recognition (ELISALF) [18, 20, 22]. A separate application evaluated voice quality based on the acoustic voice quality index, AVQI [19], which combines, e.g., shimmer and cepstral peak prominence. For comparability, the automatic intelligibility, 0-100 (0 worst), and AVQI, 0-8 (0 best), scores were scaled linearly to fit the perceptual evaluation results from experiment 1. No automatic evaluation was obtained for the variant readings of speakers *K9S* and *UCX*.

3. Results

Results of experiment 1 and the automatic evaluation scores were recalculated to pairwise differences between T2 and T1 (score at T2 minus score at T1). The four pairwise result scores of each speaker in Experiment 2 were averaged to a single preference score between [-500, 500] (after subtraction of 500). This procedure averages out any T1/T2 order bias. The averaging was done with the two remaining scores for the one speaker with a missing pair (*UCX*).

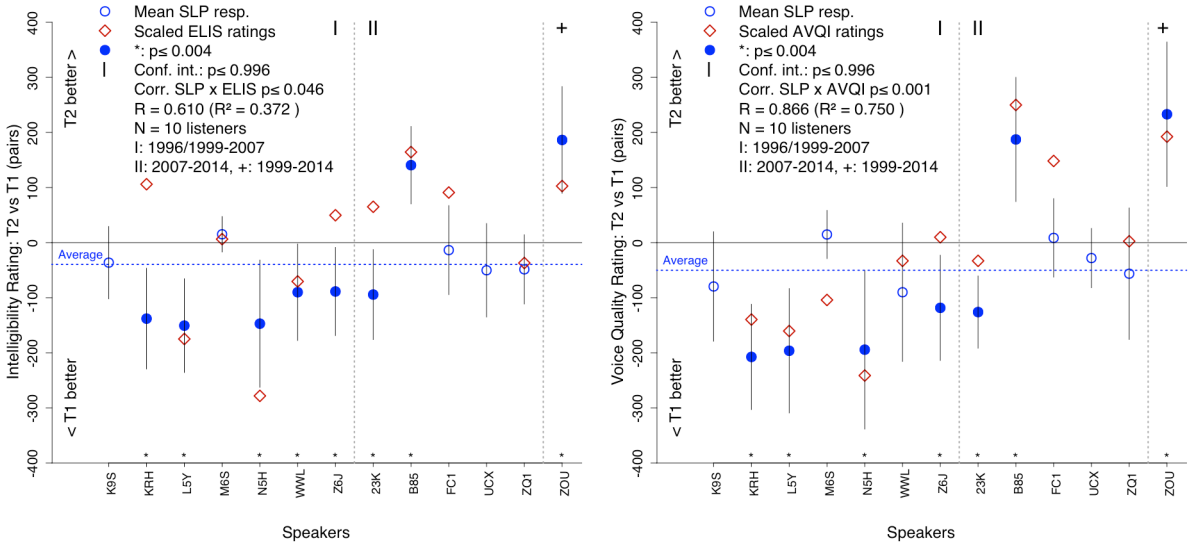

Figure 2: *Pairwise comparisons, experts and ASISTO ratings. Left: Intelligibility and ELIS, Right: Voice quality and AVQI. Statistics based on Student t-test and Pearson's product-moment correlation.*

The use of different recording equipment and procedures can introduce a bias in the evaluations. A test comparing the T1 results between periods I and II and the T2 results between periods II and III (experiment 1) showed that averaged ratings between these periods were not different (Student t-test, p>0.05). However, the small number of speakers makes the power of these tests low (c.f. Table 1). To determine for which of the 13 speakers the evaluations differed, a level of significance of p≤0.004 is used (Bonferroni correction). Statistical tests were performed in R [23].

Experiment 1

The variation in the perceptual scores in experiment 1 was high (see Figure 1). Only two speakers had statistically significant lower perceptual *intelligibility* scores for T2 than T1 (p≤0.004, not shown). The ELIS and ELISALF scores were strongly correlated with pooled perceptual *intelligibility* scores (R>0.80, p<0.001, n=24) and for T1 and T2 separately (R>0.78, p<0.005, n=12 each).

The perceptual *voice quality* scores differed for three speakers, two speakers had lower scores for T2 than T1 and one had higher scores (p≤0.004). For all other speakers, the differences were not statistically significant (p>0.004). The AVQI scores were moderately correlated with pooled perceptual *voice quality* scores (|R|>0.60, p<0.002, n=24) and for T1 separately (|R|= 0.70, p<0.02, n=12), but not for T2 (|R|=0.45, p>0.05, n=12). Perceptual *intelligibility* and *voice quality* T2-T1 difference scores were strongly correlated (R=0.89, p<0.001, n=13). ELIS and AVQI T2-T1 differences were also correlated (R=0.75, p<0.01, n=11).

The consistency of the evaluations was estimated by correlating the scores of individual experts against the average score of all the other experts (n=9). The correlations were between *R*=0.6 and *R*=0.9 for both Intelligibility and Voice Quality. Automatic scores were correlated with the average of all ten experts. The consistency of the scores for ELIS and ELISALF compared favorably against the *intelligibility* scores of individual experts: *R*≥0.8. Correlation of automatic AVQI scores was comparable to the least consistent expert: *R*≈0.6.

Experiment 2

Eight speakers showed a statistical significant difference in *intelligibility* between T1 and T2 in experiment 2 and seven of them also showed a difference in *voice quality* (see Figure 2). The ELIS speaker difference scores were modestly correlated with the average pairwise perceptual ratings for intelligibility (R=0.61, p≤0.05). The correlation of the ELISALF difference scores with the perceptual ratings was even marginally lower (R=0.58, p>0.05). The correlation between ELIS and ELISALF difference scores was statistically not significant (R=0.56, p>0.05). Because of this, we focus on the ELIS scores for the remainder of this paper. *Intelligibility* and *voice quality* were strongly correlated (R=0.99, p<0.001). AVQI scores were strongly correlated to *voice quality* and thus also to *intelligibility* (R=0.87 and R=0.84, p≤0.001). This makes AVQI a better predictor of perceptual *intelligibility* in experiment 2 than the ELIS scores. Differences between periods I and II in Figure 2 were not significant for ELIS or AVQI (p>0.025, Bonferroni correction).

Overall, five speakers had statistically significant worse *intelligibility* at T2 (T2-T1 < 0), two speakers were better at T2 (T2-T1 > 0), and four were neither better nor worse (T2-T1 ~ 0), see Figure 2. One speaker was scored with worse *intelligibility* at T2 and unchanged *voice quality* (*WWL*). In total, there are roughly as many speakers that showed a decline in *intelligibility* and *voice quality* at T2 as showed unchanged or improved *intelligibility* and *voice quality*. The ELIS scores tended to score the T2 as more intelligible than the T1 recordings. Currently, it is not clear how to interpret this difference with perceptual *intelligibility* scores. The AVQI scores were distributed more like the corresponding perceptual *voice quality* scores, in line with the high correlation between AVQI and *voice quality* scores.

In this sample of 13 speakers, three distinct levels of change can be distinguished: better at T2, worse at T2, and no difference. When speakers from each of these levels are compared against speakers from other levels a statistical significant difference is found ($p < 0.001$).

Together, the automatic and perceptual results presented in Figure 2 indicate that there is no definite trend in the changes in *intelligibility* and *voice quality* after 7 years or more. There might be a somewhat bigger probability for a decline in *intelligibility* and *voice quality* than the reverse. However, it is clear that the differences between speakers in direction and extend of change over time are large.

Consistency between recordings

For one speaker, *KRH*, there were three evaluated recordings over a span of 18 years, one from each recording period. All three recordings were used to get a rough (N=1) estimate of the variability in evaluation outcomes (Table 3, Bonferroni correction $p \leq 0.01$). It appears that the experts can judge the speech samples quite consistently. Only the *voice quality* results for period I in experiment 1 differed from the other periods (I versus II and III, $p<0.01$). None of the other evaluations differed between periods ($p>0.01$). Pairwise comparisons showed significant differences in experiment 2 ($p<0.004$, underlined), except for voice quality between periods II-III. The automatic scores for this speaker, ELIS and AVQI, were rather stable over this time course (Table 3, Experiment 1), but the difference scores were variable (Table 3, Experiment 2).

4. Discussion

Long-term stability of voice quality and intelligibility in TE-speakers, to our knowledge, has not yet been described. This study presents a unique dataset, in which perceptual and automatic voice assessment complement each other. It must be noted, though, that speech samples of only a small group of speakers was available. Differences in surgical techniques and treatment modalities are not included in this study because of the small sample size. The voice recordings were made in three time periods, and different audio recording equipment was used each time (Table 2). Our analysis did not reveal any systematic differences between time periods that could be attributed to these equipment differences. For future research, we are collecting recordings in a consistent setting.

The anatomical and physiological changes in voice production, which TL patients are facing, are immense. In TE-speech, voice is produced by the PES that originally does not have a function in sound production. Some TE-speakers present a fairly good voice, whilst others are rated as more deviant in *voice quality* and *intelligibility*. The differences between recordings vary. On average, a slight decrease over time is seen in perceptually rated *voice quality* and *intelligibility* (Figure 2). This might indicate an effect of aging.

Table 3. *Results in Experiment 1 and 2 for speaker KRH. Intell. : Intelligibility, VQ: Voice Quality.* *: *p<0.01 with other periods. _: p<0.004. See text.*

	Experiment 1			Exp. 2	
Period	I	II	III	I-II	II-III
Intell.	801	739	731	**-138**	**-66**
ELIS	620	726	581	106	-145
VQ	*690	443	461	**-208**	-97
AVQI	474	334	409	-140	75

The perceptual evaluations tend to be scattered between the expert raters (Figure 1). In the literature it is stated that expert raters such as SLP's provide more reliable outcomes than naïve listeners. To assess the consistency of the raters, for one speaker three recordings were evaluated. It appears that the experts can judge the speech quite consistently (Table 3). Using pairwise comparisons, as in experiment 2, is more sensitive to differences. Pairwise comparison results in more consistent ratings than rating individual samples, as in experiment 1.

Changes in *voice quality* and *intelligibility* are dependable within individual speakers. When *voice quality* is rated as good by perceptual evaluation, *intelligibility* tends to be as well. The strong correlation ($R=0.99$, $p<0.001$ in experiment 2) between these outcome measures confirms this dependency. The fact that independent automatic measures, AVQI and ELIS, are also correlated shows that this correlation is part of the speech signal itself. These (high) correlations indicate that *intelligibility* problems with TE substitute voices might emerge from a lower perceptual *voice quality*.

The AVQI was developed for analyzing a combination of sustained vowels and running speech samples [19]. There were no sustained vowel recordings for some of our speakers. Therefore, AVQI analysis was partially performed, i.e., on running speech only. Our results show that this procedure already provided sufficient information (c.f. [20]). The AVQI scores correlate strongly with *voice quality* scores, and therefore also with intelligibility. Since perceptual *voice quality* and *intelligibility* are strongly correlated it is shown that for these speakers, AVQI provides consistent information on both perceived *voice quality* and *intelligibility*. The AVQI was an even better predictor of perceived *intelligibility* than the automatic ELIS scores.

Ideally an automatic speech analysis program, which detects differences over time, is needed. The ELIS evaluation tool is used to evaluate individual speech samples. Comparisons between (T2 – T1) samples are made afterwards. For the future it would be recommended to develop an automatic assessment tool that can directly evaluate differences between speech samples.

5. Conclusions

Voice quality and intelligibility of TE-speakers is more or less stable over a period of 7 to 18 years. There might be a slight decrease in the quality of the TE-speech in some speakers, but, if at all present, this could not be consistently ascertained. *Voice quality* and *intelligibility* are correlated when rated perceptually by experts as well as when evaluated automatically. To get more insight in the long-term changes of speech quality it is recommended to systematically collect data of a large group of TE-speakers over a longer period of time. Tools for automatic evaluation of speech quality are very promising for analyzing trends within individual speakers.

6. Acknowledgements

The Department of Head and Neck Oncology and Surgery of the Netherlands Cancer Institute received an unrestricted research grant from Atos Medical (Hörby, Sweden). The authors have no other funding, financial relationships, or conflicts of interest to disclose.

7. References

[1] C. J. van As-Brooks, F. J. Koopmans-van Beinum, L. C. Pols, and F. J. Hilgers, "Acoustic signal typing for evaluation of voice quality in tracheoesophageal speech," *Journal of Voice,* vol. 20, pp. 355-368, 2006.

[2] E. C. Ward and C. J. van As-Brooks, *Head and neck cancer: treatment, rehabilitation, and outcomes*: Plural Publishing, 2014.

[3] F. J. Hilgers, A. H. Ackerstaff, M. van Rossum, I. Jacobi, A. J. Balm, I. B. Tan*, et al.*, "Clinical phase I/feasibility study of the next generation indwelling Provox voice prosthesis (Provox Vega)," *Acta oto-laryngologica,* vol. 130, pp. 511-519, 2010.

[4] P. Farrand and R. Endacott, "Speech Determines Quality of Life Following Total Laryngectomy: The Emperors New Voice?," in *Handbook of Disease Burdens and Quality of Life Measures*, ed: Springer, 2010, pp. 1989-2001.

[5] E. D. Blom, M. I. Singer, and R. C. Hamaker, "Tracheostoma valve for postlaryngectomy voice rehabilitation," *Annals of Otology, Rhinology & Laryngology,* vol. 91, pp. 576-578, 1982.

[6] S. Singer, M. Merbach, A. Dietz, and R. Schwarz, "Psychosocial determinants of successful voice rehabilitation after laryngectomy," *Journal of the Chinese Medical Association,* vol. 70, pp. 407-423, 2007.

[7] B. O. de Coul, F. Hilgers, A. Balm, I. Tan, F. Van den Hoogen, and H. Van Tinteren, "A decade of postlaryngectomy vocal rehabilitation in 318 patients: a single Institution's experience with consistent application of provox indwelling voice prostheses," *Archives of Otolaryngology–Head & Neck Surgery,* vol. 126, pp. 1320-1328, 2000.

[8] E. Lundström and B. Hammarberg, "Speech and voice after laryngectomy: perceptual and acoustical analyses of tracheoesophageal speech related to voice handicap index," *Folia Phoniatrica et Logopaedica,* vol. 63, pp. 98-108, 2011.

[9] M. J. McAuliffe, E. C. Ward, L. Bassett, and K. Perkins, "Functional speech outcomes after laryngectomy and pharyngolaryngectomy," *Archives of Otolaryngology–Head & Neck Surgery,* vol. 126, pp. 705-709, 2000.

[10] W. M. Mendenhall, C. G. Morris, S. P. Stringer, R. J. Amdur, R. W. Hinerman, D. B. Villaret*, et al.*, "Voice rehabilitation after total laryngectomy and postoperative radiation therapy," *Journal of Clinical Oncology,* vol. 20, pp. 2500-2505, 2002.

[11] C. J. van As, F. J. Koopmans-van Beinum, L. C. Pols, and F. J. Hilgers, "Perceptual evaluation of tracheoesophageal speech by naive and experienced judges through the use of semantic differential scales," *Journal of speech, language, and hearing research,* vol. 46, pp. 947-959, 2003.

[12] M. F. Ramírez, F. G. Doménech, S. B. Durbán, M. C. Llatas, E. E. Ferriol, and R. L. Martínez, "Surgical voice restoration after total laryngectomy: long-term results," *European archives of oto-rhino-laryngology,* vol. 258, pp. 463-466, 2001.

[13] P. H. Dejonckere, P. Bradley, P. Clemente, G. Cornut, L. Crevier-Buchman, G. Friedrich*, et al.*, "A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques," *European Archives of Oto-rhino-laryngology,* vol. 258, pp. 77-82, 2001.

[14] M. Moerman, J.-P. Martens, and P. Dejonckere, "Multidimensional assessment of strongly irregular voices such as in substitution voicing and spasmodic dysphonia: A compilation of own research," *Logopedics Phoniatrics Vocology,* vol. 40, pp. 24-29, 2015.

[15] M. Moerman, J.-P. Martens, M. Van der Borgt, M. Peleman, M. Gillis, and P. Dejonckere, "Perceptual evaluation of substitution voices: development and evaluation of the (I) INFVo rating scale," *European Archives of Oto-Rhino-Laryngology and Head & Neck,* vol. 263, pp. 183-187, 2006.

[16] M. Moerman, G. Pieters, J.-P. Martens, M.-J. Van der Borgt, and P. Dejonckere, "Objective evaluation of the quality of substitution voices," *European Archives of Oto-Rhino-Laryngology and Head & Neck,* vol. 261, pp. 541-547, 2004.

[17] T. Most, Y. Tobin, and R. C. Mimran, "Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production," *Journal of communication disorders,* vol. 33, pp. 165-181, 2000.

[18] R. Clapham, C. Middag, F. Hilgers, J.-P. Martens, M. Van Den Brekel, and R. Van Son, "Developing automatic articulation, phonation and accent assessment techniques for speakers treated for advanced head and neck cancer," *Speech Communication,* vol. 59, pp. 44-54, 2014.

[19] Y. Maryn, M. De Bodt, and N. Roy, "The Acoustic Voice Quality Index: toward improved treatment outcomes assessment in voice disorders," *Journal of communication disorders,* vol. 43, pp. 161-174, 2010.

[20] R. P. Clapham, J.-P. Martens, R. J. van Son, F. J. Hilgers, M. M. van den Brekel, and C. Middag, "Computing scores of voice quality and speech intelligibility in tracheoesophageal speech for speech stimuli of varying lengths," *Computer Speech & Language,* vol. 37, pp. 1-10, 2016.

[21] J.-P. Martens. (2016). *ASISTO*. Available: https://asisto.elis.ugent.be/

[22] C. Middag, R. Clapham, R. Van Son, and J.-P. Martens, "Robust automatic intelligibility assessment techniques evaluated on speakers treated for head and neck cancer," *Computer speech & language,* vol. 28, pp. 467-482, 2014.

[23] R. C. Team, "R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012," ed: ISBN 3-900051-07-0, 2014.