

Tracheoesophageal Speech: a Dedicated Objective Acoustic Assessment

Thomas Drugman^{a,*}, Myriam Rijckaert^b, Claire Janssens^b, Marc Remacle^b

^a*TCTS Lab, University of Mons, Mons, Belgium*

^b*Otolaryngology Service, Mont-Godinne Hospital, University of Louvain, Belgium*

Abstract

After total laryngectomy, the placement of a tracheoesophageal (TE) prosthesis offers the possibility to recover a new voice. However, the quality of the resulting TE speech is known to be degraded. To assess a patient's voice, current approaches rely either on quality-of-life questionnaires or on a perceptual evaluation carried out by speech therapists. These two methods exhibit the disadvantage of being both subjective and time-consuming. In this paper, we propose a dedicated scale, called A4S, for the objective Automatic Acoustic Assessment of Alaryngeal Speech. For this purpose, we first identify the artefacts existing in TE speech. These are linked to the periodicity, regularity, high-frequency noise and gargling noise/creakiness of the signal, as well as to the speaking rate. Specific acoustic features are proposed for the characterization of each artefact. A statistical study shows that TE speakers have a significantly worse voice compared to the control group, except for the speaking rate. Based on these advances, the A4S scale is proposed. This scale is made of five normalized dimensions, related to the five identified artefacts. A given patient's phonation can then be represented by a pentagon in a radar chart, which allows a fast and intuitive visualization of the strengths and flaws of the voice. A4S can then be seen as a useful tool for speech therapists to design tailored exercises specific to the patient's voice. In addition, we show the applicability of A4S for the follow-up of patients, as well as to study the impact of the type of surgery

*Corresponding author. Tel. +3265374749.

Email addresses: thomas.drugman@umons.ac.be (Thomas Drugman),
myriam.rijckaert@uclouvain.be (Myriam Rijckaert), janssens.claire@gmail.com
(Claire Janssens), marc.remacle@uclouvain.be (Marc Remacle)

(open neck, robot and flap reconstruction) used for total laryngectomy and of a pre-surgical radiotherapy on various aspects of the TE voice.

Keywords: Total Laryngectomy, Tracheoesophageal Speech, Voice Rehabilitation, Voice Assessment, Speech Analysis

1. Introduction

A Total Laryngectomy (TL) refers to the removal of the entire larynx and separation of the airway from the mouth, nose and esophagus. As a consequence, patients who underwent TL cannot produce speech sounds in a conventional manner because their vocal folds have been removed. The main goal of the post-surgery process therefore consists in regaining a new voice. Three main options are nowadays possible for voice restoration after TL: esophageal, electrolaryngeal and tracheoesophageal speech. In this manuscript, we focus on the analysis of tracheoesophageal (TE) speech as it has been shown in several studies (Baggs and Pine, 1983; Most et al., 2000; Robbins et al., 1984; Singer et al., 2012), to lead to superior voice rehabilitation capabilities compared to both esophageal and electrolaryngeal speech.

Since the esophagus and trachea are separated after TL, a hole called *tracheostoma* is created in the patient's neck to allow breathing. In TE speech, a surgical fistula (called *TE puncture*) is created in the wall separating the trachea and esophagus, allowing the placement of a phonatory prosthesis. The physiological differences before laryngectomy, and after the creation of the TE puncture are illustrated in Figure 1. The TE phonatory prosthesis acts like a one-way valve: an airflow can pass from the trachea to the esophagus and further to the vocal tract cavities, allowing the production of speech; the passage from the esophagus to the trachea is however made impossible to avoid that food or drinks ingested by the patient penetrate in the trachea and go down to the lungs.

When producing TE speech, the airflow passing through the phonatory prosthesis generates, for some patients, the vibration of some residual organs called the *pharyngoesophageal (PE) segment*. When a control of this *neovibrator* (also sometimes referred to as *neoglottis*) is possible, patients are able to produce voiced sounds, but generally with a lower level of periodicity. As a consequence, albeit TE speech allows to recover a new means to communicate, it suffers most of the time from a clear diminution of naturalness and

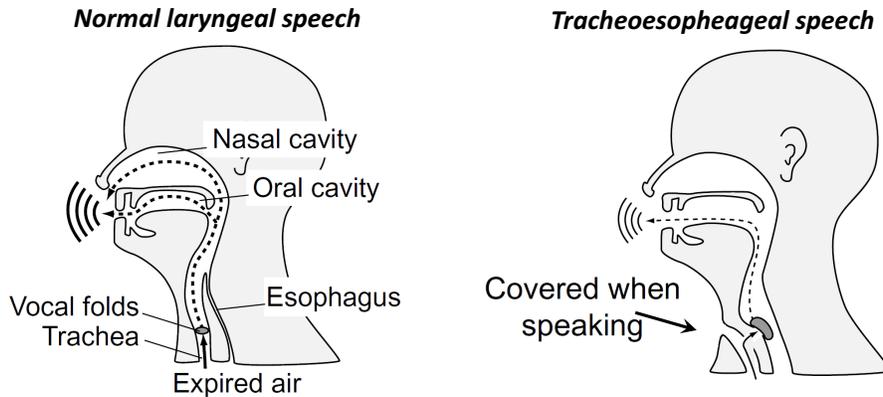


Figure 1: *Physiology of the vocal apparatus (adapted from (Nakamura, 2010)) for: Left panel: a normal subject with laryngeal speech; Right panel: a patient with a TE prosthesis. In this latter case, the tracheostoma must be covered (either with a finger or using a valve) when speaking.*

intelligibility (Most et al., 2000). In addition, the individuality/personality of the speaker is often lost (especially for female patients), mostly because of two main reasons: *i*) the vibration of the neoglottis generally occurs at a (much) lower fundamental frequency compared to the normal laryngeal voice; *ii*) the spectral shaping imposed by the vibration of the neoglottis can be radically different from that of the vocal folds. In (Singer et al., 2012), these drawbacks were observed even in a more pronounced way for esophageal and electrolaryngeal speech .

Several works have targeted the evaluation of the way TE speech is perceived. Its acceptability and intelligibility have been compared in (Most et al., 2000) to those of both laryngeal and esophageal speech. Although both aspects are degraded when compared to normal laryngeal speech, it turns out that TE speech is perceived to be more acceptable than good esophageal speech while they have a similar level of intelligibility. The study conducted in (Singer et al., 2012) investigated the evolution of the intelligibility in alaryngeal speech during the first year following the TL surgery. Patients with a TE prosthesis were observed to have the best results. The authors also emphasized the improvement made by these patients within the first year, as well as the necessity to attend rehabilitation sessions during that period. Several works have targeted the evaluation of the way TE speech is perceived. Its acceptability and intelligibility have been compared in (Most

et al., 2000) to those of both laryngeal and esophageal speech. Although both aspects are degraded when compared to normal laryngeal speech, it turns out that TE speech is perceived to be more acceptable than good esophageal speech while they have a similar level of intelligibility. The study conducted in (Singer et al., 2012) investigated the evolution of the intelligibility in alaryngeal speech during the first year following the TL surgery. Patients with a TE prosthesis were observed to have the best results. The authors also emphasized the improvement made by these patients within the first year, as well as the necessity to attend rehabilitation sessions during that period.

The comparison between TE speech and laryngeal or other alaryngeal voices has also been addressed based on some acoustic parameters. In (Robbins et al., 1984), TE speech has been analyzed using frequency, intensity and duration features. Based on these characteristics, TE speech was shown to be more similar to normal speech than is esophageal speech, and to be more intense than both other types of speech. Most et al. (2000) observed that most of the differences between laryngeal and alaryngeal speech lies in the fundamental frequency of the signal. An acoustic signal typing system based on a visual inspection of a narrow-band spectrogram was proposed in (Van As-Brooks et al., 2006). According to this visualization tool, the user can classify TE speech from a given patient into one of four pre-defined categories. The authors also showed the link of this classification with some acoustic features (standard deviation of F0, jitter, proportion of voiced speech and the band energy difference). The study conducted in (Siric et al., 2012) investigated the differences between TE and esophageal speech based on the following acoustic measurements: intensity, maximum phonation time, F0, jitter, shimmer, and Harmonic-to-Noise Ratio (HNR). The authors did not find any statistically significant differences and concluded that every patient requires an individual approach. However, the average value of the F0, the intensity and the maximum phonation time demonstrated the superiority of TE speech rehabilitation.

In the great majority of cases, TE speech is assessed using questionnaires completed by the patient and/or a dedicated perceptual scale used by the speech therapist. Several patient self-assessment instruments have been designed to measure quality of life specific to dysphonia. These include the Voice Handicap Index (VHI, (Moerman et al., 2004)), which measures the psycho-social handicapping effects of voice disorders, the Voice-Related Quality of Life Measure (V-RQOL, (Schindler et al., 2012)) or the Euro-

pean Organization of Research and Treatment of Cancer (EORTC, (Alvarez-Buylla Blanco and Herranz Gonzalez-Botas, 2011)) general questionnaires. Perceptual evaluation involves a rating of the phonation made by speech therapists using a dedicated scale. For this purpose, the *GRBAS* scale (De Bodt et al., 1997) has been designed for the general assessment of voice disorders. The *IINFVo* scale was proposed in (Moerman et al., 2006) specifically for the assessment of substitution voicing. In this scale five parameters are defined: overall impression (I), impression of intelligibility (I), unintended additive noise (N), fluency (F) and voicing (Vo). Finally, the work in (De Bodt et al., 2002) aimed at expressing the intelligibility of dysarthric speech as a linear combination of voice quality, articulation, nasality and prosody.

Although the aforementioned instruments allow an assessment of TE speech, they however exhibit several drawbacks: *i*) they are subjective as they rely on the judgement of the patient himself, or of the speech therapist. Resulting measurements will then be affected by an inherent variability; *ii*) they are relatively time-consuming. In order to overcome these drawbacks, the goal of this paper is to develop automatic acoustic tools allowing an objective and quantitative assessment of TE speech. This manuscript builds upon our preliminary study described in (Drugman et al., 2013b). Compared to (Drugman et al., 2013b), we here further extend the study of the artefacts present in TE speech to reach five relevant dimensions, for the characterization of which automatic specific acoustic tools are now developed and compared with other state-of-the-art techniques. These five relevant dimensions include the periodicity and the regularity of the signal, its high-frequency noise content, the speaking rate and the amount of creakiness/gargling in the voice. Based on the proposed dimensions, the differences between normal laryngeal (including a control set of elderly subjects) and TE speech are studied on a larger panel of subjects (containing 63 recordings of TE speech, and 21 of control subjects). Furthermore, a statistical study is carried out to investigate the significance of these differences. These new tools are embedded in a proposed scale dedicated to the objective assessment of TE speech, and called *A4S* (standing for Automatic Acoustic Assessment of Alaryngeal Speech). Finally, we illustrate the applicability of *A4S* in the follow-up of patients and in studying the impact of the type of surgery used for TL, and we discuss its perspectives.

The outcome of this work is two-fold. First, the proposed scale allows an objective assessment of the quality of the patient’s voice through several dimensions. This information can be used by speech therapists/laryngologists

for various purposes: *i*) to focus on specific aspects of the voice (as highlighted by the proposed assessment), *ii*) to compare various voice rehabilitation approaches, *iii*) to keep a follow-up of the patient. Secondly, the knowledge of these artefacts is essential when designing speaking aid systems aiming at resynthesizing an enhanced version of TE speech (as in (Qi et al., 1995) or (Del Pozo and Young, 2006)). Indeed, in order to improve the naturalness and intelligibility of TE speech, developed methods have to integrate procedures to alleviate such artefacts.

As stated above, some studies in the literature have already reported an acoustic analysis of TE speech (Most et al., 2000), (Robbins et al., 1984), (Van As-Brooks et al., 2006), (Siric et al., 2012). In addition to these, the work described in (Huang et al., 2009) investigated the use of existing acoustic analysis techniques, originally developed for the analysis of narrowband telephone speech, in order to estimate the TE speech quality. Results showed that a subset of the features from the International Telecommunication Union recommendation P.563 algorithm estimate the TE speech quality with better correlation than previously proposed features (Huang et al., 2009). With the same goal in mind, the Moore-Glasberg auditory model was employed in (McDonald et al., 2010) to extract perceptually relevant features from the acoustic signal. Results showed that the correlation between the subjective ratings and the objective predictions obtained using the features from the auditory model was better than with the features commonly used in the telecommunication industry. The Moore-Glasberg auditory model however requires the use of a reference signal (i.e. speech for the same sentence to serve as a baseline, usually of good quality), and consequently of some time-frequency mapping between the tested and reference signals. Finally the approach in (Maier et al., 2009) proposed to predict subjective scores based on the word recognition rate (WRR) obtained by an Automatic Speech Recognition system. It was shown that WRR strongly correlates with the intelligibility of the voice. A similar automatic speech recognition-based assessment of TE speech was carried out in Haderlein (2007).

Aforementioned existing studies however generally suffer from several drawbacks which we try to overcome in this paper. First, possible artefacts have never been categorized and the assessment of TE speech based on an acoustic analysis generally involves a rating along one single dimension: the perceptual voice quality or the intelligibility. Secondly, the acoustic analysis either requires a manual inspection of signals or is based on available automatic tools generally used in a *black box* manner. These latter tools have

generally been designed for normal laryngeal speech, have a low robustness and are therefore not suited for the analysis of TE speech. Furthermore, most of the measures are derived from the F0 information whose estimation is problematic if the analysis tools are inappropriate. Third, studies generally either involve a limited number of TE patients, or are only based on sustained vowels. In this paper, we target an automatic analysis conducted on read continuous speech from a rather large number of patients with a TE puncture. Artefacts are categorized and robust automatic methods for their acoustic characterization are developed and integrated in the proposed *A4S* scale.

The paper is structured as follows. Section 2 gives a description of the database used throughout our study. The artefacts present in TE speech are investigated in Section 3 where automatic analysis methods are developed for their characterization. The proposed *A4S* scale is described in Section 4 and its potential applicability is demonstrated in Section 5, where its perspectives are also discussed. Finally, Section 6 concludes the paper and discusses our future works.

2. Database

The corpus we used throughout our experiments consists of three sets: *TTS databases*, *Control* and *TE*. In the first set, we considered recordings collected at the Language Technologies Institute at Carnegie Mellon University with the goal of developing unit selection Text-To-Speech (TTS) synthesizers. More precisely, we used data from 7 speakers (5M, 2F) of the CMU ARCTIC corpus¹, with 30 utterances per speaker. This set is used as a reference of normophonic high-quality voices recorded in studio conditions. The two other datasets were acquired by speech therapists at the Mont-Godinne hospital with a high-quality handheld recorder (Olympus LS-5) with an external lapel microphone (Olympus ME-52W) designed for noise cancellation. Subjects were asked to read a phonetically-balanced text of 10 sentences (whose average duration is about 40 seconds for normophonic subjects). The *TE* set consists of recordings from 36 patients (28M, 8F) having undergone a total laryngectomy and with a TE puncture. Some of these patients have been followed-up across several sessions spaced by several weeks, leading to

¹CMU ARCTIC speech synthesis databases, http://festvox.org/cmu_arctic/

a total of 63 recordings of TE speech. They are aged between 50.2 and 84.3 years (mean = 67.9). The time elapsed between the prosthesis placement and recordings varies between 2 weeks and 10 years (mean = 641 days). 18 out of these 36 patients have undergone a traditional open neck surgical technique (i.e. by external cervical means), 6 have undergone a surgery with flap reconstruction, while a robot surgery was used for 12 of them. In the *Control* set, speech therapists recorded 14 speakers (6M, 8F) from a similar age (range= 51-72, mean=60 years) who never suffered from any voice pathology. They are here used as a comparison point for the *TE* set (same recording conditions, same age). Note that the data from the 3 sets were downsampled to 16 kHz for analysis.

3. Analysis of the Acoustic Artefacts in Tracheoesophageal Speech

After a careful listening and inspection of our 63 recordings of TE speech, we identified five main types of artefacts in TE speech related to the periodicity, the regularity, the high-frequency noise, the gargling noise/creakiness of the speech signal, and to the speaking rate. Some of them are linked to the dimensions used in the GRBAS or IINFVo perceptual rating scales (e.g. unintended additive noise, fluency and voicing (Moerman et al., 2006)). In this section, these artefacts are analyzed and quantified based on an automatic acoustic study. Since a reliable automatic estimation of the voiced segments in TE speech is yet an unsolved issue, our approach is driven as follows:

- The analysis is performed on segments with speech activity, regardless of a voicing criterion. These segments are identified as those with a total perceptual loudness (Peeters, 2003) exceeding by more than 25 dB the minimum loudness in the utterance.
- The proposed features have to be robust and independent of F0 as much as possible.
- To avoid the possible detrimental effects due to some spurious estimated values, each phonation is characterized by the median of the extracted acoustic features.

Relying on this methodology, the five identified artefacts are now analyzed and quantified, allowing a comparison of TE speech and normal laryngeal speech along these five dimensions. For each artefact, the statistical

significance of the differences between the *TE* and *Control* groups is studied. For this, the hypothesis that data come from a normally distributed population is first tested in each group using the Lilliefors test (Lilliefors, 1967). If positive, Fisher’s F-test for the null hypothesis that these two normal populations have the same variance is carried out. If these two tests are both positive, then the conditions to perform Student’s t-test for independent groups are met. Otherwise, a non-parametric Mann-Whitney-Wilcoxon (MWW, (Mann and Whitney, 1947)) test is carried out. For both tests, the Bonferonni correction is applied.

3.1. Periodicity of the Speech Signal

TE speech has been observed in the literature to be less periodic, with pitch values comparable to those in normal speech (Van As-Brooks et al., 2006). Nonetheless, these results were obtained either from a manual input with a visual inspection of spectrograms, or from an automatic analysis using the Praat toolkit (Boersma, 1993) as a *black box*. However the two pitch tracking methods available in Praat (using either the autocorrelation or cross-correlation function) are known to have a poor robustness (Drugman and Alwan, 2011). It is therefore not surprising to find spurious F0 values up to more than 400 Hz (Siric et al., 2012), which is completely unrealistic in TE speech. As a consequence, the resulting F0-derived measures are sometimes suspicious and should be considered cautiously.

In this work, the periodicity analysis relies on the Summation of the Residual Harmonics (SRH, (Drugman and Alwan, 2011)) algorithm which was especially designed for robust pitch tracking. SRH first estimates the spectral envelope by conventional linear prediction analysis and then inverse-filters the speech signal. This step allows to considerably remove the influence of the vocal tract resonances and of the noise. The spectrum of the resulting residual excitation signal is then investigated and the SRH criterion exploits its harmonic structure. SRH then not only estimates the F0 track but also the level of periodicity/voicing in the speech signal. A reliable binary voicing decision can then be taken by applying a simple threshold (fixed to 0.07 as in (Drugman and Alwan, 2011)) to SRH values. Based on this criterion, we define the voicing proportion as the percentage, among the frames with speech activity, of frames detected as voiced.

Another feature capturing the periodicity of the source excitation is the kurtosis of the Linear Prediction (LP) residual signal (Drugman, 2014; Falk et al., 2012; Drugman, 2013). In unvoiced speech, the excitation signal has

a noisy nature and its kurtosis is relatively low (Drugman, 2014), while in voiced speech, it exhibits quasi-periodic discontinuities at the Glottal Closure Instants (GCIs, (Drugman et al., 2012b)) which is reflected by a greater sparsity and hence higher values of the kurtosis (Drugman, 2014). The kurtosis of the LP residual signal was used in (Falk et al., 2012) as an efficient way to characterize the vocal harshness in spastic dysarthric speech.

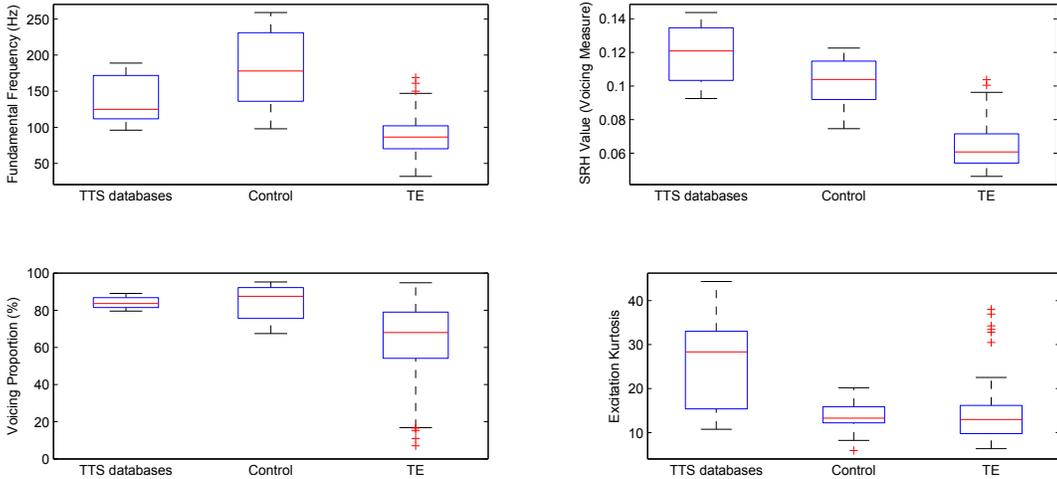


Figure 2: *Distributions of the periodicity measures across the three datasets. Top left panel: the fundamental frequency F_0 ; Top right panel: the SRH values indicating the degree of voicing; Bottom left panel: the proportion of voiced speech ; Bottom right panel: the kurtosis of the excitation signal.*

The distributions of these 4 periodicity-based measures for the 3 datasets are presented in Figure 2 in the form of boxplots. It can be observed that for the F_0 in TE patients, the third quartile is at about 100 Hz. In other words, only 25% of the TE patients have a pitch comparable to what is produced in normal laryngeal speech, while others have much lower pitch values. Interestingly, this observation holds for both genders. It is worth noting that the average pitch was 84.2 Hz for female TE patients, against 93.7 Hz for male TE voices. This emphasizes the importance of developing systems aiming to enhance TE speech, as this loss of identity is difficult to live with, particularly for female patients. A second conclusion from Figure 2 is that TE speech is much less periodic compared to normal speech, with SRH values significantly lower. This is reflected in the voicing proportion which is reduced for the majority of TE patients. More precisely, 6 out of the 36

patients were almost unable to produce any voicing and communicated nearly uniquely with whispered speech. An important remark from the periodicity study is the large discrepancy amongst TE patients, not only regarding their ability to generate voiced sounds, but also with regard to their mean pitch. For example, while one male patient produced pitch contours around 30 Hz, another used F0 values at about 160 Hz. Finally, it can be observed that the periodicity in the *Control* set was found to be lower than in the *TTS* set. This decrease is due to aging, as known from the literature (Dehgan and Scherer, 2013). Note that all these results derived from the use of the SRH algorithm have been corroborated by a visual inspection of the signal. In other words, the speech waveform and its spectrogram have been checked manually to ensure that the SRH results are relevant and valid for the analysis of TE speech, which was confirmed through our observations.

The statistical study between the *TE* and *Control* groups indicated that they are significantly different ($p < 0.001$ using the MWW test) for the 3 considered SRH-based features. Regarding the kurtosis of the excitation signal, it can be seen that *Control* and *TE* groups have much lower values compared to the *TTS* dataset. However, the differences between the *Control* and *TE* voices did not turn out to be significant. One possible explanation is the fact that elderly normophonic voices are known to already exhibit a certain harshness (Dehgan and Scherer, 2013).

3.2. Regularity of Phonation

In addition to the reduced periodicity, we observed the TE phonation to be less regular. This can be physiologically explained by the fact that turbulences are more important at the pharyngoesophageal (PE) segment for TE patients, than at the glottis for normal subjects. The amount of irregularities is here assessed based on four acoustic measures: the variation of the Chirp Group Delay (CGD), the spectral variation, the normalized LP error and the Short-Term Temporal Dynamics. The variation of the CGD was proposed in (Drugman et al., 2011b) as an efficient measure to characterize voice disorders. The advantage of using the CDG is that this representation is based on the phase of the speech signal, which was shown to be more suited for capturing its irregularities compared to its Fourier amplitude counterpart (Drugman et al., 2011b). The spectral variation (Peeters, 2003) is computed as the normalized cross-correlation between two successive amplitude spectra (using 30 ms-long Hanning windows, and a conventional frame shift of 10 ms). The more irregular the signal, the greater the CGD and spectral

variations. The normalized LP error is the error made when considering an auto-regressive (AR) model (whose order is standardly fixed to $F_s/1000 + 2$, where F_s denotes the sampling rate) to explain the speech signal. It is defined as the prediction error divided by the energy of the frame. If the speech production satisfied the ideal AR model, voiced speech would be characterized by a LP residual signal being an ideal pulse train, and the LP error would be minimum. The stronger the turbulences during the phonation, the more the excitation signal contains noise and irregularities, and the more it deviates from the ideal pulse train. An irregular phonation is thus reflected by larger normalized LP errors. Finally, the Short-Term Temporal Dynamics (STTD) was proposed in (Falk et al., 2012) to characterize rhythm pattern perturbations. STTD can be computed as the standard deviation of the log-energy rate of change, defined as the first derivative of zeroth order cepstral coefficient. It was shown in (Falk et al., 2012) to be a good indicator of dysarthric word intelligibility.

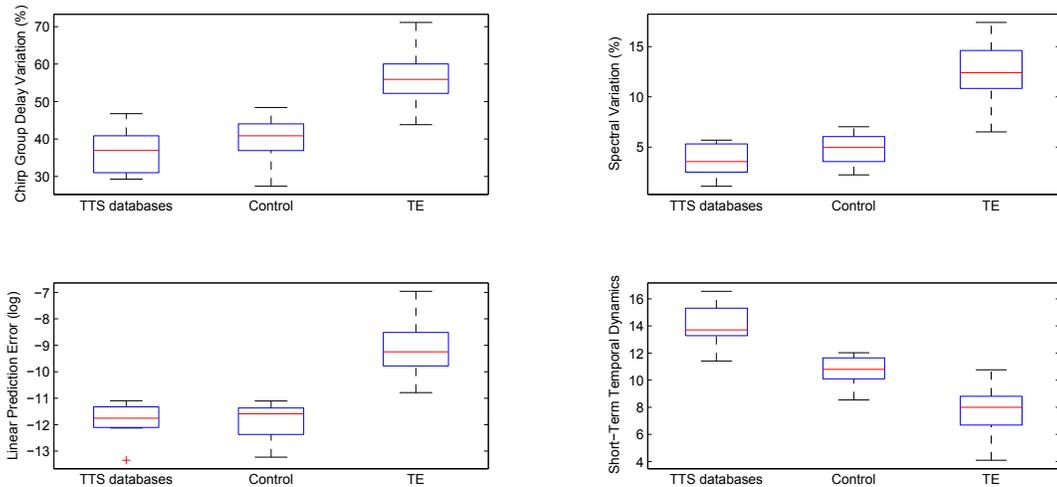


Figure 3: *Distributions of the regularity measures across the three datasets. Top left panel: the Chirp Group Delay variation; Top right panel: the spectral variation; Bottom left panel: the normalized linear prediction error; Bottom right panel: the Short-Term Temporal Dynamics.*

Results are shown in Figure 3. These distributions reflect the same phenomenon: the regularity in the *TE* group is much lower. These differences were observed to be statistically significant ($p < 0.001$) for the four parameters (employing a t-test for the CGD, spectral variations and STTD, and a

MWW test for the normalized LP error). Again, the effect of aging can be seen in the *Control* set which is less regular compared to the *TTS* databases. It is worth noting at this stage that periodicity and regularity are two complementary aspects of speech. As it will be discussed in Section 4.1, these two dimensions are only poorly correlated. For example, we noted that some patients are able to produce TE speech with an acceptable periodicity but with a low regularity, or vice versa.

3.3. High-Frequency Noise

Some patients encounter difficulties in managing the airflow arising in the trachea such that it completely goes through the TE prosthesis when speaking. That can be the case when the patient uses a valve at the tracheostoma. As a consequence, a certain airflow is evicted at the tracheostoma which not only leads to a less loud speech, but also to the impression of a noisier signal, particularly in the high frequencies (HF). In order to quantify the amount of HF, the long-term average spectrum is estimated for each speaker. For this, the amplitude spectrum of each frame (for which speech activity has been detected) is computed and normalized in energy. Obtained spectra are then averaged for the whole phonation. Since the text to be read is phonetically balanced, the effects of formants can be assumed to cancel each other out, and the long-term spectrum contains averaged contributions of the vocal tract and of the (either glottal or neoglottal) source. The HF noise is finally measured as the relative energy beyond a given frequency (fixed to 1.5 kHz in this work) in the long-term spectrum.

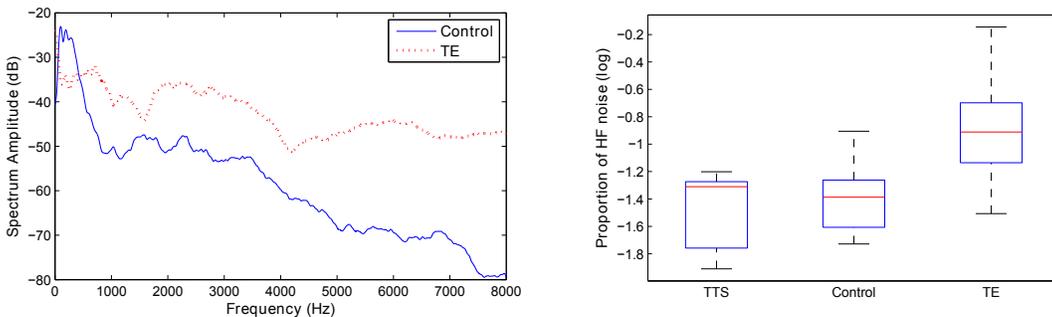


Figure 4: Left panel: *Example of long-term averaged spectra for a standard control subject and a TE patient whose speech is considerably affected by uncontrolled noise*; Right panel: *distribution of the proportion of HF noise (beyond 1.5 kHz) for the 3 datasets.*

An example of a long-term spectrum is displayed in the left panel of Figure 4 for both a standard control subject and a TE patient who produces a considerable amount of HF noise. The box plot of the proportion of HF noise is given in the right panel of Fig. 4. It can be seen that, on average, most of the TE patients have a greater amount of high-frequencies in their speech. The Student’s t-test showed these differences to be statistically significant ($p < 0.001$). Some patients with a TE prosthesis are however able to produce speech with a proportion of HF similar to that of normal speech. On the opposite, for a few others, the amount of HF noise can be relatively high. One way to explain these differences is the fact that the production at the PE segment can be strongly altered compared to the vibration at the glottis in normal laryngeal speech. As a consequence, TE speech lacks the spectral shaping imposed by the glottal source and is mainly characterized by a low-frequency resonance called glottal formant (Drugman et al., 2011a), (Drugman, 2011). As stated above, this effect might be further reinforced by a lack of control of the valve at the tracheostoma.

3.4. Gargling Noise/Creakiness

For some patients, speech is perceived as if they were talking with water in their throat. This is typically due to deglutition problems, which lead to the fact that saliva and/or nasal mucus may flow down in the throat. Because of these secretions, the resulting speech signal may sporadically exhibit artefacts, as illustrated in Figure 5 for a vowel /a/. The smoothed Hilbert envelope is indicated to highlight the slow-varying fluctuations of energy. It can be observed that the gargling noise is reflected by uncontrolled energy bursts in the signal (generally spaced by more than 30 ms). The resulting perceptual effect is usually close to that of a creaky voice. Generally speaking, the sensation of a creaky voice quality is often found in TE speech, while its usage is relatively rare in normal laryngeal speech. Creaky voice is defined as a raspy or croaking quality of the voice generally produced with a very low pitch and often with highly irregular periodicity (Laver, 1980). It is generally interpreted based on the auditory criterion “a rough quality with the additional sensation of repeating impulses”, as is done in (Ishi et al., 2008a), (Kane et al., 2013). The creakiness dimension is then somehow linked to the *Roughness* axis in the GRBAS perceptual scale (De Bodt et al., 1997).

In order to automatically detect the regions of creaky voice in continuous speech, we use the algorithm first proposed in (Drugman et al., 2012a) and further developed in (Kane et al., 2013). This method is based on acoustic

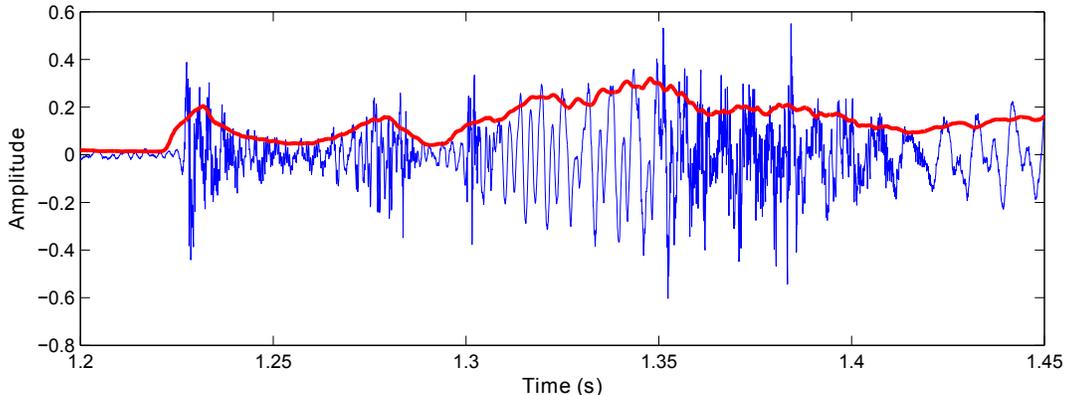


Figure 5: *Example of a TE speech segment exhibiting gargling noise, which is reflected by uncontrolled energy bursts.*

features specifically designed for the characterization of creaky voice, and which are the input of an artificial neural network (ANN). The acoustic features used for this purpose were proposed in (Drugman et al., 2012a) and (Ishi et al., 2008b), and include: *i*) the difference $H2 - H1$ in dB between the first two harmonics of the output of a resonator whose center frequency is set to the speaker’s mean F0, *ii*) the residual peak prominence, which characterizes the prominence of LP-residual peaks relative to its immediate neighborhood, *iii*) power peak parameters derived from the very short-term power contour, *iv*) inter-pulse similarity, *v*) intra-frame periodicity. Based on these features, an ANN classifier was trained on several hours of manually-annotated speech in several languages and communicative contexts. The algorithm finally gives a binary decision at the frame level estimating whether or not speech has been creaked. In this paper, the amount of gargling noise and creakiness (as they are reflected through similar perceptual properties) is quantified as the proportion of frames detected as creaky using the aforementioned algorithm.

Another feature characterizing the gargling noise present in the voice is the reverberation-to-signal modulation energy ratio (RSMR), originally introduced in (Falk et al., 2010). This feature exploits the modulation spectral characteristics of clean speech to compare the modulation energy between the signal and the “*room reverberation*”. In the context of the original feature definition, an additional noise was found to be caused by room reverberation (Falk et al., 2010). For TE speech, it was shown in (Huang et al., 2009) that these additional modulation frequencies are due to artefacts (e.g. gargling

noise, raspiness) found in the voice. In (Huang et al., 2009), RSMR was observed to be the most efficient feature for the automatic estimation of the TE speech quality.

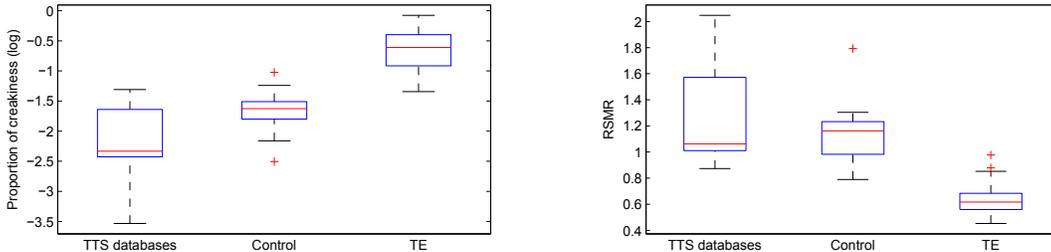


Figure 6: *Distributions of the creakiness/gargling measures across the three datasets. Left panel: the proportion of creakiness; Right panel: the reverberation-to-signal modulation energy ratio.*

Figure 6 displays, for the three datasets, the distributions of the proportion of creakiness (on a logarithmic scale) and of RSMR. It turns out that speakers in the *TTS* and *Control* sets have a creaky usage never exceeding 10%. This goes in line with the findings in (Drugman et al., 2013a) drawn over a range of languages (US English, Japanese, Swedish and Finnish) where creaky voice was used between 3.5 and 10.5% of the time (as extracted from manual annotations). The artefacts in *TE* speech however lead to a perception of creakiness largely exceeding 10% for about 3 patients over 4. This rate goes even up to 68% in the worst case. It can be also seen that *TE* voices are characterized by lower RSMR values. After statistical analysis, these differences between the *TE* and *Control* groups turned out to be significant for both features ($p < 0.001$ with a MWW test).

3.5. Speaking Rate

It is known that some *TE* patients face some issues in maintaining a certain fluency when reading a text. They generally have to make more or longer breaks to allow saliva/mucus swallowing or breath taking. This has been reflected by taking into account a *Fluency* parameter in the perceptual IINFVo scale (Moerman et al., 2006). This disfluency is characterized by a reduced speaking rate, here estimated as the ratio of frames for which speech activity has been detected (according to the perceptual loudness criterion given in the beginning of Section 3). Note that silences at the beginning and

end of the recording (before and after the speaker is actually reading) have been obviously discarded.

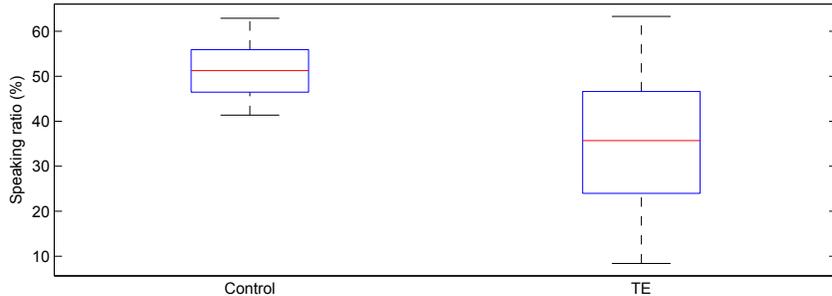


Figure 7: *Distribution of the speaking ratio for the Control and TE datasets.*

The distribution of the speaking ratio is shown in Figure 7 for the *Control* and *TE* datasets. It is worth noticing that the results obtained for the *TTS* databases are not included here, as they are not comparable to those of the *Control* and *TE* datasets. Indeed, the task in which the speakers were involved is different. In the *TTS* set, speakers were asked to utter each sentence separately and the database segmentation was carried out accordingly. Silences in between sentences have therefore been removed, and computing the speaking ratio as defined above would not make sense here. On the contrary, subjects from the *Control* and *TE* datasets were asked to read the same text continuously, and they freely imposed their pausing time following their abilities. Consequently we here compare only the speaking ratio of these two latter sets. It can be observed that only 50% of the *TE* patients were able to read the text with a speaking rate comparable to what is achieved by the control subjects. The rest of them encountered difficulties in sustaining a sufficient speaking rate and had to use more frequent and/or longer pauses. The Student’s t-test showed that speaking rate for *TE* patients was not significantly lower after Bonferonni correction ($p = 0.0223 > 0.05/9$).

4. The Proposed A4S Scale

In Section 3, we have identified five acoustic artefacts which may occur in *TE* speech. A thorough acoustic and statistical analysis has been performed and tools have been developed in order to automatically quantify

these artifacts. In this section we derive a scale for the Automatic Acoustic Assessment of Alaryngeal Speech, called A4S. The proposed scale is now described in Sections 4.1, and the correlation between its axes is studied in Section 4.2.

4.1. The A4S Scale

The A4S scale allows a fast, objective and automatic evaluation of substitution voices. It builds upon the acoustic analysis tools developed in Section 3 and is here evaluated in the frame of TE speech assessment. The A4S scale consists of an evaluation of the voice along the five following dimensions: i) *periodicity*, measured via the voicing proportion; ii) *regularity*, measured via chirp group delay variation; iii) *HF noise*, measured via the relative energy beyond 1.5 kHz in the long-term spectrum; iv) *gargling noise/creakiness*, measured via the proportion of speech frames detected as creaky voice; v) *speaking rate*, measured via the proportion frames with a speech activity. These five axes have been normalized between 0 and 1, such that 0 corresponds to the worst value in the *TE* dataset and 1 to the best value in the *Control* group. The A4S space is then represented under the form of a radar chart in which a given voice is then characterized by a specific pentagon.

Figure 8 illustrates the resulting A4S scale where five voices are represented. The averaged TE and Control voices are indicated for information purpose, so as to more easily compare the evaluated voice with reference patterns. The three other voices are related to the phonation of a given patient 2 weeks, 3 weeks and 3 months after the placement of his TE Provox prosthesis. This patient is 58 years old, has undergone a traditional open neck surgery and did not follow any complementary surgical or radiotherapy treatment at the time of the recordings. A4S allows the speech therapist or the laryngologist to rapidly figure out what are the flaws and strengths of the voice, and to appreciate the evolution of the patient’s voice. In the specific case of the patient studied in Figure 8, it turns out that he initially suffered from severely lower values along the speaking rate and gargling/creakiness axes, while his voicing and production of HF noise were largely above the averaged TE patient (and comparable to the averaged control subject). As the patient attended the rehabilitation sessions, his voice got better which is reflected by a general increase of the pentagon area in the A4S plane. For this particular case, it can be notably observed that the patient made considerable improvements by alleviating the flaws and consolidating the strengths of his voice.

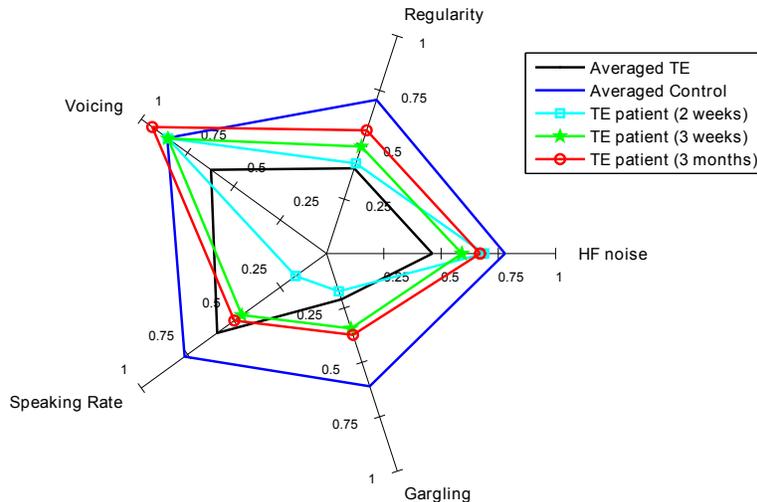


Figure 8: *Illustration of the A4S scale for a 58-year old patient recorded 2 weeks, 3 weeks and 3 months after the placement of the TE prosthesis.*

4.2. Correlation between the A4S axes

Ideally, the acoustic axes of the A4S scale should be independent such that they characterize different perceptual dimensions of the voice. Table 1 shows the Pearson correlation coefficient ρ between the five axes of the A4S scale, computed for the 63 recordings of the *TE* dataset. It turns out that values are relatively weak, except for the periodicity and HF noise axes which seem to be relatively highly correlated ($\rho = 0.785$). This reflects the fact that the production of voicing by TE patients is possible thanks to the PE segment which imposes a low-frequency spectral shaping, with most of the energy in the first harmonics. On the opposite, for patients who are almost unable to produce any voicing, there is not such a low-frequency vibration of the PE segment and the resulting speech signal contains relatively less energy below 1.5 kHz. Although the correlation between the periodicity and HF noise dimensions is non-negligible, we chose to keep them as distinct axes in the A4S scale as it is important to laryngologists and speech therapists to be able to quantify these two aspects of the voice separately.

All other correlation measures in Table 1 are rather low ($\rho < 0.35$), indicating a good independence between the axes. Finally, let us notice the negative correlation between the HF noise and the gargling noise/creakiness. As stated in Section 3.4, the gargling noise and voice creakiness are char-

Table 1: Pearson correlation coefficient between the five axes of the A4S scale.

	Per.	Reg.	HF noise	Gargl.	Sp. Rate
Periodicity	1	0.303	0.785	0.238	0.221
Regularity	.	1	0.348	0.140	0.332
HF noise	.	.	1	-0.208	0.316
Gargling/Creaky	.	.	.	1	-0.003
Speaking Rate	1

acterized by repeating impulses occurring at a very low frequency. A voice exhibiting such an artefact will then inherently comprise relatively less high frequencies, and therefore have lower values for the proportion of HF noise.

5. Applicability of the A4S scale

The applicability of the proposed A4S method will be here shown for two specific tasks: *i*) comparing the type of surgery used for total laryngectomy, *ii*) following-up patients. Experiments run for these two applications are described in Sections 5.1 and 5.2, and the potential perspectives of A4S are summarized in Section 5.3.

5.1. Influence of the type of surgery and of a pre-surgical radiotherapy on the voice quality

For this experiment, the A4S scores along the five axes were extracted for the 63 recordings of our TE speech database. These scores were then analyzed according to the type of surgery used for TL. Results are summarized in Figure 9. One first interesting observation is that patients with an *open neck* surgery (i.e. by external cervical means) had, on average, a better score than subjects with a flap reconstruction or with a robot surgery, and this along the five A4S axes. For the comparison with the flap reconstruction, these differences are significant only along the *Regularity* axis ($p = 0.0031$), while they are significant along the *HF noise* ($p < 0.001$), *Regularity* ($p = 0.0049$) and *Periodicity* ($p < 0.001$) axes for the comparison with the robot surgery. The greatest difference lies in the ability to reproduce voicing, where patients with an open neck surgery got an averaged periodicity of 0.747, against only 0.407 for those having undergone a robot surgery.

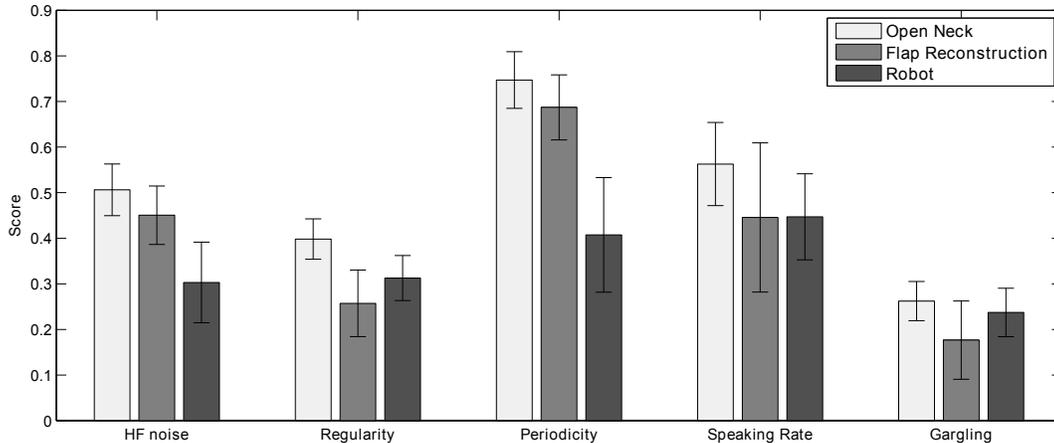


Figure 9: Comparison of the techniques used for TL surgery based on the scores along the five A4S axes, together with their 95% confidence intervals.

The fact that the voice quality for patients who underwent a robot surgery is much worse than that of TE subjects who underwent an open neck surgery can be explained as follows. The majority of our *Robot* patients had to be treated by radiotherapy and/or chemotherapy prior to the total laryngectomy surgery. The quality of their pharyngeal tissue was consequently affected. In such cases, a traditional open neck surgery by external cervical means is most of the time avoided as it generally leads to complications (necrosis, fistula, vascular rupture, etc.). For these patients, using the robot is the most appropriate type of surgery as it guarantees the respect of the skin and of the pre-laryngeal tissues, and engenders much simpler post-surgical after-effects. However, the pharyngeal tissues remain fibrous, which makes the vibration of the pharyngoesophageal (PE) segment more difficult. Furthermore, for these same patients, a salivary bypass tube is generally placed for an averaged duration of 3 months. When removing this latter tube, the opening of the pharyngeal tract is wider than for patients who underwent an open neck surgery. This further reduces the ability of the PE segment to vibrate properly. It is therefore not surprising to observe in Figure 9 a much lower periodicity for the patients from the *Robot* dataset.

In order to further investigate the impact of a pre-surgical radiotherapy and/or chemotherapy on the TE voice quality, we have split the *Robot* group into the patients who had to follow radiotherapy and/or chemotherapy in

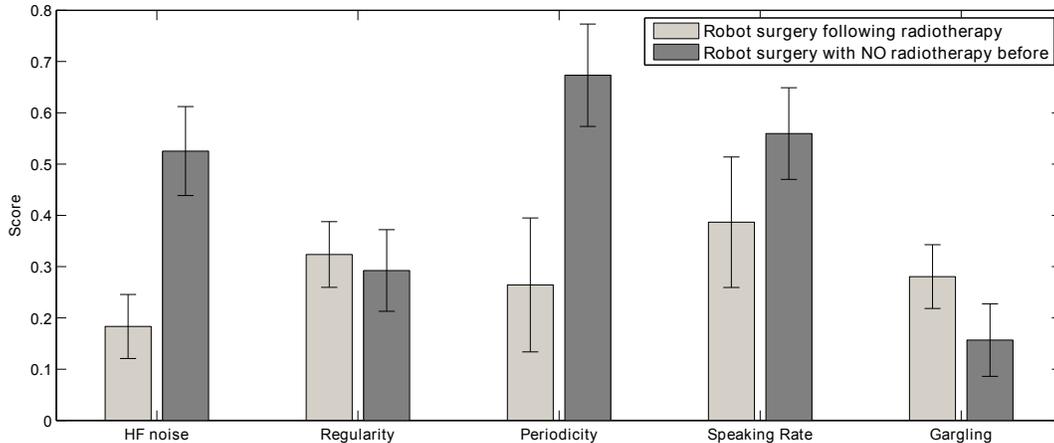


Figure 10: *Impact on the A4S axes of radiotherapy and/or chemotherapy in the month preceding the TL, for the patients who underwent a robot surgery.*

the month preceding the TL, and those who did not. Results are displayed in Figure 10. Patients who did not have radiotherapy and/or chemotherapy prior to the surgery are clearly able to reproduce voicing in a significantly better way ($p = 0.0043$) and to better control the amount of HF noise ($p < 0.001$). This is because the mucus membrane of their neoglottis is softer, which therefore allows a better vibration. Finally, note that on overall the voice of patients with no radiotherapy before the TL exhibits more gargling noise ($p = 0.0265$).

5.2. Patient Follow-Up

For this second experiment, we have conducted a follow-up for 12 patients. Medical information about these subjects can be found in Table 2. Half of them have undergone an *open neck* surgery, the robot surgery was used for 5 of them, while only one underwent a total larygectomy with flap reconstruction. Two did not follow any complementary treatment, six had only radiotherapy, while four of them had to follow both radiotherapy and surgical treatments.

Each patient was followed-up and was recorded in several sessions after the placement of his/her TE prosthesis. For each recording, the acoustic features detailed in Section 3 were automatically extracted and the scores for the five axes of the A4S scale were calculated as described in Section 4. Each recording was finally characterized by its total A4S score averaged across

Table 2: Medical information for the 12 patients who have been followed-up.

	Gender	Age	Surgery	Compl. Treatment
Patient 1	M	69	Open Neck	None
Patient 2	F	66	Open Neck	Radiotherapy
Patient 3	M	58	Open Neck	Radiotherapy
Patient 4	M	50	Open Neck	Radiotherapy
Patient 5	M	63	Open Neck	RdTh + Surgical
Patient 6	M	76	Open Neck	RdTh + Surgical
Patient 7	M	74	Robot	Radiotherapy
Patient 8	F	76	Robot	Radiotherapy
Patient 9	M	78	Robot	Radiotherapy
Patient 10	F	54	Robot	RdTh + Surgical
Patient 11	F	66	Robot	RdTh + Surgical
Patient 12	M	65	Flap Reconstr.	None

the five dimensions. Note that one could think of other ways/functions to summarize the values along the five A4S axes into a single metric. For example, a linear combination could be considered, where the weights might come from a medical advice so as to balance the preponderance of each dimension individually. A simple mean function is used throughout the remainder of this paper though.

The results of the patient follow-up are displayed in Figure 11 as a function of the time elapsed between the placement of the TE prosthesis and the recording. This procedure allows to quickly quantify the progress/regression made by the patient. Some observations can be made from the results in Figure 11: *i*) confirming the findings from Figure 9, patients who underwent an open neck surgery got on overall higher A4S scores compared to those who underwent a robot surgery; *ii*) except for Patient 7 whose voice slightly degraded, patients who underwent a robot surgery improved over time. However, for three out of the six patients who underwent an open neck surgery, a considerable regression is observed; *iii*) there does not seem to be much differences across patients receiving different types of complementary treatment. Nonetheless it is important to note that these results are limited to 12 patients, and we will avoid drawing any hasty conclusion here as these

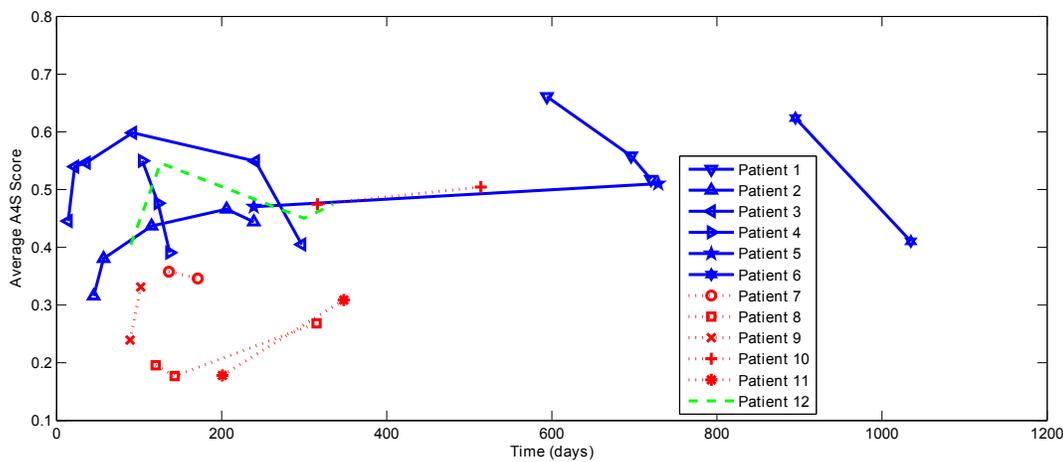


Figure 11: Results of the follow-up for 12 patients using the average A4S score. The x-axis indicates the time elapsed between the placement of the prosthesis and the recording.

observations should be confirmed in other studies.

5.3. Perspectives of the A4S scale

As highlighted by the examples shown in Figures 8, 9, 10 and 11, the perspectives of the A4S methodology are various:

- *Design of tailored speech therapy exercises:* This is the primary goal of the A4S scale. As A4S allows a quantification along its 5 axes, the strengths and flaws of the TE voice can be immediately seen on its corresponding pentagon in the A4S space (similarly to that of Figure 8). Based on this, the speech therapists can design tailored rehabilitation exercises so that the patient can work on specific aspects of his/her voice.
- *Patient follow-up:* As shown in Figure 11 or with the evolution of the pentagons in Figure 8, A4S allows to quantify the improvements made by the patient. In this way, the progress (or regression) made by the patient is clear both to the speech therapist and to the patient himself. Furthermore we believe that providing a report on the patient's voice after each speech rehabilitation session could encourage the patient to aim for better performance as this will become more challenging to

him/her (this might even be reinforced with the patient being aware of the weaknesses of his/her voice).

- *Analyzing the influence of various medical factors*: Provided that sufficient data is available to draw statistically significant conclusions, the A4S scale can be used to study the effects of the voice of various medical factors on different aspects. These factors include the type of surgery (as done in Figure 9), the complementary treatment followed by the patient or the use of either a handsfree kit or a filter-button at the tracheostoma.
- *Use of A4S to predict a non-automatic subjective assessment*: Similarly to what has been achieved in (Huang et al., 2009) or (McDonald et al., 2010), the acoustic features proposed in this manuscript could be used to predict subjective ratings of the TE voice quality. Furthermore, since A4S characterizes various aspects of the voice, it could be interesting to investigate the correlation of its axes with perceptual dimensions such as those used in the *GRBAS* (De Bodt et al., 1997) or in the *IINFVo* (Moerman et al., 2006) scales.

6. Conclusions and Future works

The goal of this paper was to propose a dedicated scale, called A4S, for the objective Automatic Acoustic Assessment of Alaryngeal Speech. A4S is derived from a study of the artefacts existing in TE speech. This study identified five main artefacts for which specific acoustic features were developed for their automatic characterization. These artefacts include a lower periodicity and regularity, a greater amount of HF noise, the presence of gargling noise and creakiness in the speech signal, and a reduced speaking rate. A statistical analysis showed that the differences between TE and normal laryngeal speech are significant, except for the speaking rate. The resulting A4S scale is then made of five dimensions related to these artefacts, for which one relevant measure was chosen. The A4S scale is presented under the form of a radar chart, allowing a fast and intuitive visualization of the patient's voice, and an easy determination of its strengths and flaws. The applicability of A4S was finally illustrated by comparing the types of surgery used for total laryngectomy, and by conducting a patient follow-up.

A4S is currently used by the speech therapists and laryngologists of the Mont-Godinne hospital in Belgium for the objective evaluation of the pa-

tient's voice, as a complement to standard quality-of-life questionnaires. Our future goals encompass: *i*) using the acoustic features proposed in this paper to predict the perceptual ratings used in the GRBAS and IINFVo scales, or to estimate the subjective speech quality rated by normal hearing listeners, *ii*) investigating the role played by various medical factors on the quality of the TE voice.

7. Acknowledgements

Thomas Drugman is supported by FNRS.

References

- Alvarez-Buylla Blanco, M., Herranz Gonzalez-Botas, J., 2011. Quality of life evolution in patients after surgical treatment of laryngeal, hypopharyngeal or oropharyngeal carcinoma. *Acta Otorrinolaringol Esp.* 62, 103–112.
- Baggs, T., Pine, S., 1983. Acoustic characteristics: tracheoesophageal speech. *J Commun Disord.* 16, 299–307.
- Boersma, P., 1993. Accurate short-term analysis of fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proc. Inst. Phonetic Sci.* 17, 97–110.
- De Bodt, M., Hernandez-Diaz, H., Van De Heyning, P., 2002. Intelligibility as a linear combination of dimensions in dysarthric speech. *Journal of Communication Disorders* 35, 283–292.
- De Bodt, M., Wuyts, F., Van de Heyning, P., Croux, C., 1997. Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *Journal of Voice* 11, 74–80.
- Dehgan, A., Scherer, R., e. a., 2013. The Effects of Aging on Acoustic Parameters of Voice. *Folia Phoniatr Logop.* 64, 265–270.
- Del Pozo, A., Young, S., 2006. Continuous tracheoesophageal speech repair. *Proc. European Signal Processing Conference.*
- Drugman, T., 2011. *Advances in Glottal Analysis and its Applications.* PhD Thesis, University of Mons, Belgium.

- Drugman, T., 2013. Residual Excitation Skewness for Automatic Speech Polarity Detection. *IEEE Signal Processing Letters* 20, 387–390.
- Drugman, T., 2014. Maximum Phase Modeling for Sparse Linear Prediction of Speech. *IEEE Signal Processing Letters* 21, 185–189.
- Drugman, T., Alwan, A., 2011. Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics. *Interspeech*.
- Drugman, T., Bozkurt, B., Dutoit, T., 2011a. Causal-anticausal Decomposition of Speech using Complex Cepstrum for Glottal Source Estimation. *Speech Communication* 53, 855–866.
- Drugman, T., Dubuisson, T., Dutoit, T., 2011b. Phase-based information for voice pathology detection. *IEEE ICASSP*, 4612–4615.
- Drugman, T., Kane, J., Gobl, C., 2012a. Resonator-based creaky voice detection. *Interspeech*.
- Drugman, T., Kane, J., Gobl, C., 2013a. Data-driven detection and analysis of the patterns of creaky voice. *Computer Speech and Language* 28, 1233–1253.
- Drugman, T., Rijckaert, M., Lawson, G., Remacle, M., 2013b. Analysis and Quantification of Acoustic Artefacts in Tracheoesophageal Speech. *Lecture Notes in Computer Science, Advances in Non-Linear Speech Processing* 7911, 104–111.
- Drugman, T., Thomas, M., Gudnason, J., Naylor, P., Dutoit, T., 2012b. Detection of glottal closure instants from speech signals: a quantitative review. *IEEE Transactions on Audio Speech and Language processing* 20 (3), 994–1006.
- Falk, T., Chan, W., Shein, F., 2012. Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility. *Speech Communication* 54, 622–631.
- Falk, T., Zheng, C., Chan, W., 2010. A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech. *IEEE Trans. on Audio, Speech, and Language Processing* 18, 1766–1774.

- Haderlein, T., 2007. Automatic Evaluation of Tracheoesophageal Substitute Voices. Logos Verlag.
- Huang, A., Falk, T., Chan, W., Parsa, V., Doyle, P., 2009. Reference-free automatic quality assessment of tracheoesophageal speech. IEEE Conf. Eng. Med. Biol. Soc., 6210–6213.
- Ishi, C., Sakakibara, K., Ishiguro, H., Hagita, N., 2008a. A method for automatic detection of vocal fry. IEEE Transactions on Audio, Speech, and Language Processing 16 (1), 47–56.
- Ishi, C., Sakakibara, K., Ishiguro, H., Hagita, N., 2008b. A method for automatic detection of vocal fry. IEEE Transactions on Audio, Speech, and Language Processing 16 (1), 47–56.
- Kane, J., Drugman, T., Gobl, C., 2013. Improved automatic detection of creak. Computer Speech and Language 27, 1028–1047.
- Laver, J., 1980. The Phonetic Description of Voice Quality. Cambridge University Press.
- Lilliefors, H., 1967. On the KolmogorovSmirnov test for normality with mean and variance unknown. Journal of the American Statistical Association 62, 399–402.
- Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M., Nöth, E., 2009. PEAKS A system for the automatic evaluation of voice and speech disorders. Speech Communication 51, 425–437.
- Mann, H., Whitney, D., 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Annals of Mathematical Statistics 18, 50–60.
- McDonald, R., Parsa, V., Doyle, P., 2010. Objective estimation of tracheoesophageal speech ratings using an auditory model. J. Acoust. Soc. Am. 127, 1032–1041.
- Moerman, M., Martens, J., Dejonckere, P., 2004. Application of the Voice Handicap Index in 45 patients with substitution voicing after total laryngectomy. Eur Arch Otorhinolaryngol 261, 423–428.

- Moerman, M., Martens, J., Van der Borgt, M., Peleman, M., Gillis, M., Dejonckere, P., 2006. Perceptual evaluation of substitution voices: development and evaluation of the (I)INFVo rating scale. *Eur Arch Otorhinolaryngol* 263, 183–187.
- Most, T., Tobin, Y., Mimran, R., 2000. Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production. *Journal Commun. Disord.* 33, 165–180.
- Nakamura, K., 2010. Speaking-Aid Systems Using Statistical Voice Conversion for Electrolaryngeal Speech. PhD Thesis, Nara Institute of Science and Technology, Japan.
- Peeters, G., 2003. A large set of audio features for sound description (similarity and classification) in the CUIDADO project.
- Qi, Y., Weinberg, B., Bi, N., 1995. Enhancement of female esophageal and tracheoesophageal speech. *Journal of the Acoustical Society of America* 98, 2461–2465.
- Robbins, j., Fisher, H., Blom, E., Singer, M., 1984. A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production. *Journal of Speech and Hearing Disorders* 49, 202–210.
- Schindler, A., Mozzanica, F., Ginocchio, D., Invernizzi, A., Peri, A., Ottaviani, F., 2012. Voice-related quality of life in patients after total and partial laryngectomy. *Auris Nasus Larynx* 39, 77–83.
- Singer, S., Wollbruck, D., Dietz A., e. a., 2012. *Head and Neck Journal*. *Journal Commun. Disord.* doi: 10.1002/hed.23183.
- Siric, L., Sos, D., Rosso, M., Stevanovic, S., 2012. Objective assessment of tracheoesophageal and esophageal speech using acoustic analysis of voice. *Coll Antropol.* 36, 111–114.
- Van As-Brooks, C., Koopmans-van Beinum, F., Pols, L., Hilgers, F., 2006. Acoustic signal typing for evaluation of voice quality in tracheoesophageal speech. *Journal of Voice* 20, 355–368.