Analysis and Quantification of Acoustic Artefacts in Tracheoesophageal Speech

Thomas Drugman[1], Myriam Rijckaert[2], George Lawson[2], and Marc Remacle[2]

[1] TCTS Lab., University of Mons, Belgium
[2] Ontolaryngology Service, Mont-Godinne Hospital, University of Louvain, Belgium

Abstract. After total laryngectomy, the placement of a tracheoesophageal (TE) puncture offers the possibility to gain a new voice. However, the produced TE speech is known to have a lower quality and intelligibility. The goal of this paper is to identify and quantify the acoustic artefacts in TE speech. The advantage of this study is twofold. First, the proposed measures can be used by speech therapists in voice rehabilitation sessions to assess the voice of the patient, to follow up his/her evolution and to design tailored exercises. Secondly, these artefacts have to be quantified and taken into account in synthesis methods aiming at enhancing TE speech. Four categories of acoustic artefacts are identified in this work: a lower periodicity and regularity of the phonation, and the presence of high-frequency and gargling noises. Each artefact is studied and compared to normal laryngeal speech recorded either for speech synthesis purpose or by elderly people. Results quantify the importance of each of these artefacts, and show a large disparity between TE patients.

1 Introduction

Patients having undergone Total Laryngectomy (TL) cannot produce speech sounds in a conventional manner because their vocal folds have been removed. Gaining a new voice is then the major goal of the post surgery process. There are currently three options for voice restoration after TL: tracheoesophageal speech, electrolaryngeal speech and esophageal speech. In this article, we focus on the analysis of tracheoesophageal (TE) speech. Indeed, it has been shown in several studies that TE puncture leads to superior voice rehabilitation capabilities compared to the two other approaches [1], [2].

After TL, the patient's larynx has been removed and the esophagus and trachea are separated. A hole called tracheostoma is created in the patient's neck to allow breathing. In TE speech, a surgical fistula (TE puncture) is created in the wall separating the trachea and esophagus, allowing the placement of a phonatory prosthesis. Thanks to this prosthesis, an airflow passes from the trachea to the esophagus and further to the vocal tract cavities. For some patients, this airflow generates the vibration of residual organs called the pharyngoesophageal

T. Drugman and T. Dutoit (Eds.): NOLISP 2013, LNAI 7911, pp. 104–111, 2013.

(PE) segment. When patients are able control this neovibrator (also sometimes referred to as neoglottis), they can produce voiced sounds but with a lower level of periodicity. Therefore, although TE speech allows to recover a mode of communication way, it suffers from a clear diminution of naturalness and intelligibility in the produced voice. Besides, the individuality/personnality of the speaker is often lost (this is particularly true for female patients). These conclusions hold even in a more pronunced way for esophageal and electrolaryngeal speech.

The perception of TE speech has been evaluated in the literature [1], [2]. Its acceptability and intelligibility have been compared in [1] to those of both laryngeal and esophageal speech. As expected, it has been shown that both aspects are degraded with regard to laryngeal speech. Nonetheless, TE speech turns out to be more acceptable than good esophageal speech while they have a comparable level of intelligibility. In [2], Singer et al. investigated the intelligibility of alaryngeal speech during the first year after TL. It was noticed that patients with a TE puncture had the best results in intelligibility. Authors also emphasized the considerable improvement within the first year, and the importance for the patient to attend rehabilitation sessions.

TE speech has also been studied from an acoustic point of view. In [3], TE speech is analyzed using frequency, intensity and duration features. It is shown that, based on these characteristics, TE speech is more similar to normal speech than is esophageal speech, and that it is more intense than both other types of speech. The acoustic study led in [1] revealed that most of the differences between normal and laryngeal speech lies in the fundamental frequency of the speech signal. An acoustic signal typing system based on a visual inspection of a narrow-band spectrogram was proposed in [4]. According to this visualization, the user can classify TE speech from a given patient into one of four pre-defined categories. Authors also show the link of this classification with some acoustic features (standard deviation of F0, jitter, proportion of voiced speech and the band energy difference). In [5], the acoustic differences between TE and esophageal speech are studied based on the following measures: intensity, maximum phonation time, F0, jitter, shimmer, and Harmonic-to-Noise Ratio (HNR).

Finally, several approaches have targeted the resynthesis of an enhanced version of TE speech, in order to improve its quality and intelligibility. In [6], Qi et al. resynthesized female TE words with a synthetic glottal waveform and with smoothed and raised F0. It was shown that the replacement of the glottal waveform and F0 smoothing alone produced most significant enhancement, while increasing the average F0 led to less dramatic improvement. The speech repair system proposed in [7] resynthesizes TE speech using a synthetic glottal waveform, reduces its jitter and shimmer and applies a spectral smoothing and tilt correction algorithm. A subjective assessment reveals a reduction of the perceived breathiness and harshness of the voice. The solution described in [8] interestingly focuses on the speech reconstruction from whispered voice, and proposes a modified version of the CELP vocoder. Unfortunately, authors only report an improvement compared to electrolaryngeal speech, and no comparative results are given for TE speech.

The goal of this paper is to analyze and quantify the acoustic artefacts exhibited in TE speech. The applicability of this study is two-fold. First, the proposed acoustic features allow an objective assessment of the quality of the patient's voice through several dimensions. This information can be used by speech therapists for multiple 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 of paramount importance for speaking aid systems aiming at resynthesizing an enhanced version of TE speech. Indeed, in order to improve the naturalness and intelligibility of TE speech, developed methods have to integrate procedures to alleviate such artefacts.

As aforementioned, some studies in the literature have already reported an acoustic analysis/assessment of TE speech [1], [3], [4], [5]. Nonetheless, these studies generally suffer from several drawbacks which we try to overcome in this paper. First, possible artefacts have never been categorized and the assessment is generally based only on periodicity-related measures. Secondly, the acoustic analysis either requires a manual inspection of signals or is based on the use of available automatic tools in a *black box* way. 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. Besides, most of the measures are derived from the F0 information whose estimation is problematic if the analysis tools are not appropriate. Third, studies generally involve a limited number of TE patients, or are only based on sustained vowels. In this paper, we target an automatic analysis led on read speech from a sufficiently large number of patients with a TE puncture. Artefacts are categorized and robust automatic methods for their acoustic characterization are proposed.

The paper is structured as follows. First, the database used throughout all our analyses is described in Section 2. The various acoustic artefacts in TE speech are presented in Section 3. In that section, each artefact is specifically analyzed, with regard to normal laryngeal speech, and the obtained results are discussed. Finally, Section 4 concludes the paper.

2 Database

The database we used throughout our experiments consists of three sets: *TTS*, *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 corpus [9], 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 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. The *TE* set consists of recordings from 23 patients (19M, 4F) having undergone a total laryngectomy and with a TE puncture. They are aged

between 52 and 82 years (mean = 64.5). The time elapsed between the prosthesis placement and recordings varies between 3 months and 10 years (mean = 715 days). In the *Control* set, speech therapists recorded 12 speakers (6M, 6F) 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 resampled at 16 kHz.

3 Acoustic Artefacts in Tracheoesophageal Speech

After a careful listening, we identified four main types of artefacts in TE speech. In the following, 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 a yet 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 loudness exceeding by more than 25 dB the minimum loudness in the utterance.
- The extraction of acoustic information targets robust features being as independent from F0 as possible.
- To avoid the possible detrimental effects due to some spurious values, each speaker is characterized by the median of the extracted acoustic features.

The artefacts are now studied based on this methodolody.

3.1 Periodicity of the Speech Signal

The periodicity of the TE speech signal has been observed the literature to be less periodic, with pitch values comparable to those in normal speech [4]. 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 as a *black box*. However the two pitch tracking methods available in Praat (AC and CC) have been shown to have a poor robustness [10]. It is therefore not surprising to find spurious F0 values up to more than 400 Hz [5], which is irrealistic in TE speech. As a consequence, some of these results using F0-derived measures are sometimes suspicious and should be taken with caution.

In this work, the periodicity analysis relies on the Summation of the Residual Harmonics (SRH, [10]) method which was shown to clearly outperform state-of-the-art approaches for robust pitch tracking. This technique provides estimates of two periodicity characteristics: SRH values which quantify the level of periodicity in the signal, and the pitch values. As suggested in [10], the binary voicing decision is taken by applying a threshold of 0.07 to SRH values. This allows us to define the voiced proportion as the percentage of frames recognized as voiced according to this criterion. The distributions of these 3 measures for the 3 datasets are presented in Figure 1 under the form of boxplots. It is quantitatively confirmed that TE speech is much less periodic than normal speech,

with SRH values significantly lower (except for one single speaker). We noticed that patients with a TE prosthesis were able to produce voiced speech with a proportion varying from 36 to 93% (median: 77%), with the exception of two patients who almost always spoke with whispered speech. Finally, TE patients were observed to use lower fundamental frequencies (median: 99 Hz) regardless of the gender, with a large variety across patients. For example, while one male patient produced pitch contours around 30 Hz, another used F0 values at about 160 Hz. Note that these results were confirmed by a manual inspection of the signals. 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 [11].

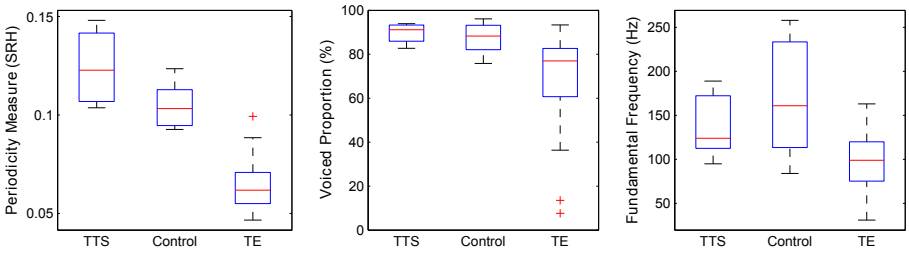

Fig. 1. *Distributions of the periodicity measures across the three datasets.* Left panel*: the SRH values indicating the level of perdiodicity,* Middle panel*: the proportion of voiced segments,* Right panel*: the fundamental frequency F0.*

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 PE segment for TE patients, than at the glottis for normal subjects. The amount of irregularities is here assessed via three acoustic measures. The first one is the variation of the Chirp Group Delay (CGD) which is a phase-based feature shown in [12] (in the frame of voice pathology detection) to be particularly suited for capturing the signal irregularities. The second is the spectral variation [13] computed as the normalized cross-correlation between two successive amplitude spectra. Finally, the third measure is the normalized Linear Prediction (LP) error, i.e. the error made when considering an autoregressive model (whose order is standardly fixed to $F_s/1000 + 2$) to explain the speech signal. If the speech production statisfied ideally this modeling, voiced speech would be characterized by a LP residual signal being an ideal pulse train, and the LP error would be minimum. The more the turbulences during the phonation, the more the excitation signal contains noise and irregularities, and the more it deviates from the ideal pulse train. This will thus be reflected in the LP error.

The distributions of these 3 acoustic measures are displayed in Figure 2. These plots reflect coherently the same phenomenon: while the regularity in *TTS* and *Control* datasets is comparable, it is observed to be significatnly lower in TE speech. This turns out to hold for all TE patients. It is worth noting at this stage that periodicity and regularity are two complementary aspects of speech. For example, the Pearson correlation coefficient between SRH and CGD values only barely reaches -0.49. In this way, we noticed that some patients can produce TE speech with an acceptable periodicity and low regularity, and vice versa.

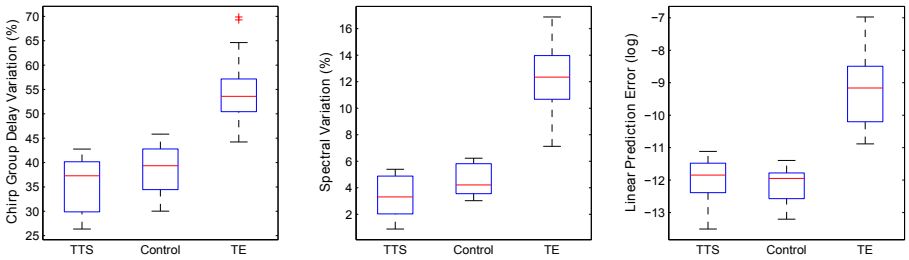

Fig. 2. *Distributions of the regularity measures across the three datasets.* Left panel*: the Chirp Group Delay variations,* Middle panel*: the spectral variations,* Right panel*: the normalized prediction error (on a logarithmic scale).*

3.3 High-Frequency Noise

For some patients, the presence of high-frequency (HF) noise can be particularly annoying. To quantify the amount of high frequencies, the long-term average spectrum is estimated for each speaker. For this, the spectrum of each frame (where speech activity has been detected) is computed and normalized in energy. Obtained spectra are then averaged over all sentences uttered by the speaker. In this way, since the text to be read is phonetically balanced, the effects of formants can be assumed to cancel each others, and the long-term spectrum contains averaged contributions of the vocal tract and of the source (either laryngeal or alaryngeal). A way to measure the average quantity of HF noise is to calculate the relative energy beyond a given frequency (fixed to 1.5 kHz in this work) in the long-term spectrum.

The left panel of Figure 3 exhibits the distributions of this measure for the 3 datasets. It can be seen that, on average, most of the TE patients have a greater amount of high-frequencies in their speech. Nonetheless, about 1 patient with a TE prosthesis over 4 produces speech with a proportion of HF similar to normal speech. On the opposite, for a few others, the amount of HF noise can be relatively high. This is illustrated in the right panel of Figure 3 where the long-term spectrum of such a TE patient is exhibited (with the one of a standard control speaker for comparison purpose). These differences can be explained by the noisy airflow evicted at the tracheostoma by some TE patients when

speaking, and by the fact that the production at the PE segment differs strongly from the vibration at the glottis in normal laryngeal speech, and consequently that the spectral shaping imposed at source is different.

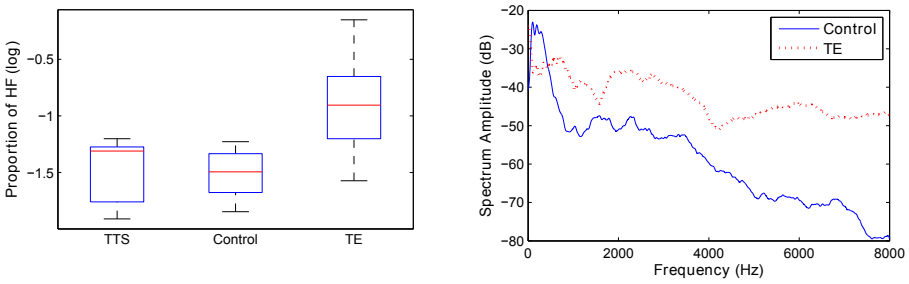

Fig. 3. Left panel: *Distribution (on a logarithmic scale) of the relative energy beyond 1.5 kHz in the long-term spectrum.* Right panel: *Examples of long-term spectra for two subjects from the* Control *and* TE *datasets respectively.*

3.4 Gargling Noise

Finally, a last artefact was observed in a minority (3 out of the 23) of patients with a TE puncture: the gargling noise. For such patients, speech is perceived as if they were talking with water in their throat. This is 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 4 for a vowel /a/. The smoothed Hilbert envelope is indicated for information purpose. It can be seen that the gargling noise is reflected by uncontrolled energy bursts in the signal (generally spaced by more than 50 ms). Note that the reliable detection and quantification of this artefact would require further investigation.

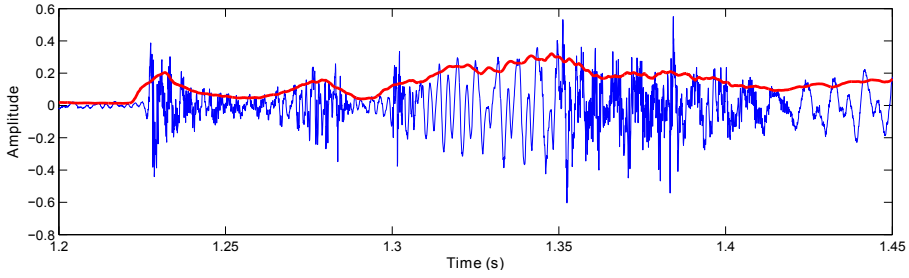

Fig. 4. Illustration of a gargling noise for a vowel /a/

4 Conclusion

This paper proposed an automatic quantification of the artefacts in tracheoesophageal speech. Four categories of acoustic artefacts were identified: a lower periodicity and regularity of the phonation, and the presence of high-frequency and gargling noises. Each artefact and its physiological origin were analyzed. Besides, robust acoustic features were proposed to characterize the first three artefacts. This allows a multidimensional assessment of the patient's voice which can be used by speech therapists during voice rehabilitation sessions. These findings are also of paramount interest for synthesis techniques targeting the enhancement of TE speech as these methods should compensated the artificats highlighted in this paper.

Acknowledgments. Thomas Drugman is supported by a FNRS Research Fellow grant.

References

1. Most, T., Tobin, Y., Mimran, R.: Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production. Journal Commun. Disord. 33(2), 165–180 (2000)
2. Singer, S., Wollbruck, D., Dietz, A., et al.: Speech rehabilitation during the first year after total laryngectomy. Head and Neck Journ. (2012) doi: 10.1002/hed.23183
3. Robbins, J., Fisher, H., Blom, E., Singer, M.: A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production. Journal of Speech and Hearing Disorders 49(2), 202–210 (1984)
4. van As-Brooks, C., Koopmans-van Beinum, F., Pols, L., Hilgers, F.: Acoustic signal typing for evaluation of voice quality in tracheoesophageal speech. Journal of Voice 20(3), 355–368 (2006)
5. Siric, L., Sos, D., Rosso, M., Stevanovic, S.: Objective assessment of tracheoesophageal and esophageal speech using acoustic analysis of voice. Coll Antropol. 36(suppl. 2), 111–114 (2012)
6. Qi, Y., Weinberg, B., Bi, N.: Enhancement of female esophageal and tracheoesophageal speech. Journal of the Acoustical Society of America 98, 2461–2465 (1995)
7. del Pozo, A., Young, S.: Continuous tracheoesophageal speech repair. In: Proc. European Signal Processing Conference, EUSIPCO (2006)
8. Reza Sharifzadeh, H., McLoughlin, I., Ahmadi, F.: Recontruction of Normal Sounding Speech for Laryngectomy Patients Through a Modified CELP Codec. IEEE Trans. on Biomedical Engineering 57(10) (2010)
9. CMU ARCTIC speech synthesis databases, `http://festvox.org/cmuarctic/`
10. Drugman, T., Alwan, A.: Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics. In: Proc. Interspeech (2011)
11. Dehgan, A., Scherer, R., et al.: The Effects of Aging on Acoustic Parameters of Voice. Folia Phoniatr Logop. 64(6), 265–270 (2013)
12. Drugman, T., Dubuisson, T., Dutoit, T.: Phase-based information for voice pathology detection. In: Proc. IEEE ICASSP, pp. 4612–4615 (2011)
13. Peeters, G.: A large set of audio features for sound description (similarity and classification) in the CUIDADO project (2003)