A Better Understanding of Esophageal Speech Excitation Source Behavior

Radhouane Bouazizi and Sofia Ben Jebara
Lab. COSIM, Ecole Supérieure des Communications de Tunis, Carthage University,
Route de Raoued 3.5 Km, Cité El Ghazala, Ariana 2083, Tunisia

Keywords: Esophageal Speech, Excitation Source, Esophagus Extremity Vibration, Opening/Closing Cycle.

Abstract: Understanding the excitation source of the esophageal speech is a key approach for understanding the esophageal speech. In this paper, we extract the excitation source using an inverse filtering approach and we analyze it. We, for example, show some similarities with an artificial EGG signal. We also detect the closing instants in order to define cycles of opening/closing of esophagus extremity and to recognize the equivalent of glottal cycles. These cycles are classified into different types according to their characteristics. A physical explanation of the esophagus extremity behavior is systematically given at the different steps of the analysis.

1 INTRODUCTION

For several reasons that range from innate, pathological (cancer for example) to accidental reasons, a person, namely called laryngectomee, may lose his voice after a laryngectomy operation (vocal cords eradication) (Brown et al., 2003). Therefore, the person becomes unable to produce speech in a normal manner. This is explained by the need of the vocal cords in the process of speech production. Esophageal voice consists in producing speech by bringing air and releasing it through the end of the esophagus which replaces the vocal cords when speaking in normal manner. One kind of esophageal speech production mechanism begins by injecting the air through the mouth in order to reach the extremity of the esophagus. While leaving, this trapped air causes the vibration of the tissue at the entrance of the esophagus. Then, as in the case of normal speech, where the role of the vocal cords is to vibrate under the action of the pressurized air ejected from the lungs, the esophagus extremity does the same thing with the incoming pressured air. The signal created at this level of speech production mechanism is called the excitation source and have a crucial role in producing substitution voice of good quality. In fact, all the remaining of the mechanism is the same as the one of normal speech (modulation in the vocal tract and radiation through the lips).

This paper aims to study in depth the excitation source of esophageal voice and the behavior of the end of the esophagus. In fact, contrary to vocal cords whose behavior is well studied and understood, the esophagus extremity behavior, as excitation source, is still not well mastered.

The paper is organized as follows. In section 2, we will show the complexity of signals (esophageal speech and its excitation source) in the temporal and frequency domains. Section 3 is devoted to establish an equivalence between an artificial ElectroGlottoGraphy signal and the source of the esophageal speech. In section 4, we define and localize the closure instants of the esophagus extremity. In section 5, different types of opening/closing cycles are identified and characterized. Finally, a conclusion is given.

2 SHOWING ESOPHAGEAL VOICE PARTICULARITY

In this section, we will first, describe the speech model that is appropriate for esophageal speech. Then, a comparison between natural and esophageal voices and their excitation sources, in time and frequency domains, is made.

2.1 Source/Filter Model of the Speech

Fig. 1 illustrates an example of the signals and the filters appearing in the source/filter model of esophageal voice. The excitation source $g(n)$ is shown in the temporal and frequency domains. It is then filtered

Bouazizi R. and Ben Jebara S..
A Better Understanding of Esophageal Speech Excitation Source Behavior.
DOI: 10.5220/0004228301630168
In *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing* (BIOSIGNALS-2013), pages 163-168
ISBN: 978-989-8565-36-5

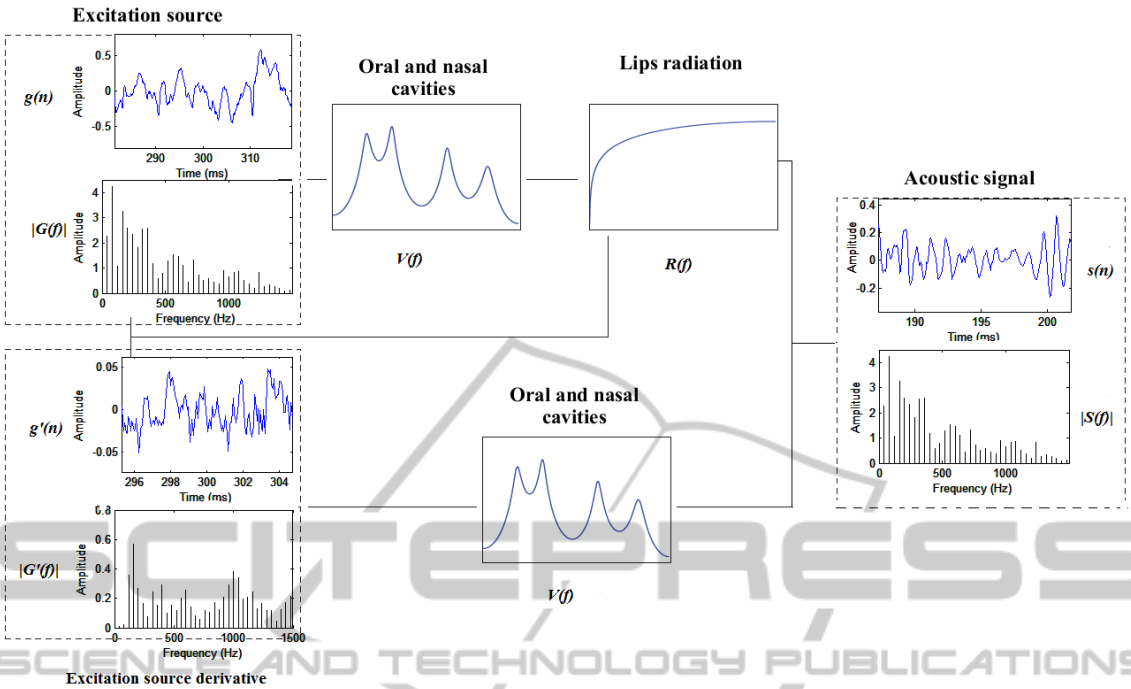

Figure 1: Illustration of source/filter model of esophageal speech.

by the oral and nasal cavities filter $V(f)$ whose frequency response is shown. Finally, the voice is produced via lips radiation whose transfer function is a simple derivative operation ($L(z) = 1 - z^{-1}$).

It is important to note that, given the linearity of the model and for simplicity reasons, one may consider that the role of the vocal tract and the lips can be inverted so that the derivative is applied immediately to the excitation source. This kind of model is illustrated in the bottom part of Fig. 1.

2.2 Time and Frequency Domains Representations of Speech

In Fig. 2.a (resp. 2.c) , an example of time domain representation of natural (resp. esophageal) speech of the phoneme /a/ is presented. As the periodicity and regularity are visually observed for natural voice, it is not the case for esophageal voice where many fluctuations that are difficult to characterize with the naked eye are observed. Globally, the signal is unstructured and the periodicity of the signal is not clear.

Fig. 2.b and 2.d show the frequency domain representation of the same signals. The natural voice is characterized by power peaks at frequencies multiples of the fundamental frequency F_0 whereas the esophageal voice does not have such structure. It is characterized by very marked peaks in the frequency region around 1 KHz.

2.3 Time and Frequency Domains Representations of the Excitation Source

To go further in the comparison between natural and esophageal speech, we propose to study the excitation source. It is estimated using the iterative adaptive inverse filtering IAIF method described in (Alku et al., 2004). It works basically as follows: the vocal tract transfer function model is iteratively estimated. Then its contribution is cancelled from the speech signal via the so-called Inverse Filtering to obtain the glottal flow of a speech signal.

In Fig. 3.a and Fig. 3.c, the glottal waves of the signals of Fig. 2 are given. The excitation source of the natural voice has a periodic shape and a well-ordered structure against a non-periodic and a miss-ordered structure of the esophageal voice excitation source.

Fig. 3.b and 3.d shows the frequency domain representations of the same signals. The one of esophageal speech is characterized by an energy concentration at low frequencies (from 0 to 600 Hz) but it is not structured in peaks placed at frequency multiples of F_0 as it is the case of the glottal flow of natural speech.

As a conclusion, since there is no clear similarity between the speech signal of the natural voice and the one of esophageal voice, the remainder of the paper

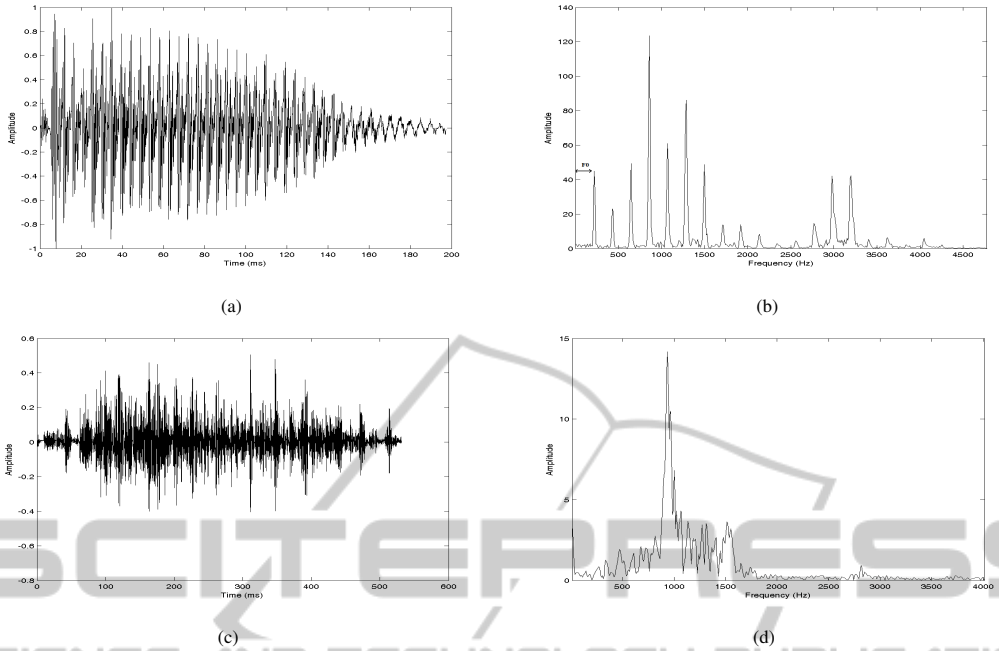

Figure 2: Time (a) and frequency (b) domains representations of the natural speech and time (c) and frequency (d) domains representations of the esophageal speech.

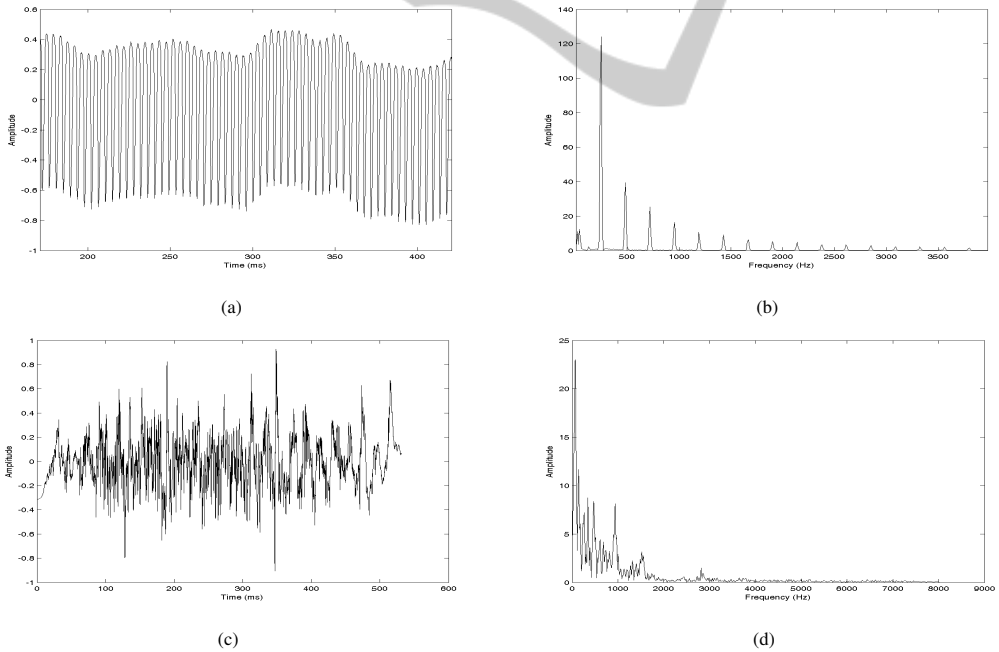

Figure 3: Time (a) and frequency (b) domains representations of the excitation source of natural speech and time (c) and frequency (d) domains representations of the excitation source of esophageal speech.

will be devoted to a deeper analysis of the excitation source. The purpose is to better understand the behavior of the esophagus extremity, the organ responsible of esophageal speech production.

3 MATCHING THE EXCITATION OF THE ESOPHAGEAL SPEECH WITH AN ARTIFICIAL EGG SIGNAL

The ElectroGlottoGraphy (EGG) is a non invasive experimental procedure used to measure the glottal flow

of natural speech. The acquired signal describes the vocal folds vibration where period cycles of opening/closing are observed. The top portion of the signal reflects opening of the vocal folds, while the bottom portion reflects closing. It can be inverted according to the electrodes connection. The EGG signal fits well the excitation source (see for example Fig.4 for the example of the natural phoneme /a/).

The question that arises is: is it possible to describe the esophagus extremity in the same manner? In other words, does the excitation source of esophageal speech looks like an EGG signal?

But, of course, in case of esophageal speech, there is no known invasive external device that helps acquiring a signal describing the behavior of the esophagus extremity in terms of opening/closing cycles.

To overcome this limitation, we propose an alternative solution, to create artificially, the equivalent of an EGG signal. This procedure is described as follows.

- Given that the main difference between these two types of voice resides in the excitation source, we propose to try to match an EGG signal generated in an artificial manner to the excitation source of the esophageal voice.
- The glottal flow of natural speech follows the well known LF model (Fant et al., 1985). So, by fixing the fundamental frequency F_0, one can generate easily an artificial EGG signal.
- The esophageal speech is not systematically periodic in all speech segments. So we try to find manually a fundamental frequency value for which the esophageal speech excitation signal fits the artificial EGG.
- The matching is done by locating areas where we can fit the maxima (resp. minima) of the EGG to the maxima (resp. minima) of the source.

Fig.5 shows the matching between the artificial EGG of a natural voice and the excitation source of the phoneme /a/ of an esophageal voice. We can see that we can locate some areas where there is an important similarity between the two signals (two examples of areas are delimited by bold discontinuous vertical lines). We can also match the position of the maxima (resp. minima) of the source to the maxima (resp. minima) of the EGG. Furthermore, we can locate, between two successive minima, a kind of period, which corresponds to a cycle of phonation duration (fundamental period). In the same figure, the two pointed periods have close values (4.3 ms and 4.15 ms). This local period allows finding a frequency that enables us to generate an artificial EGG that can fit perfectly (in a local manner) the excitation source.

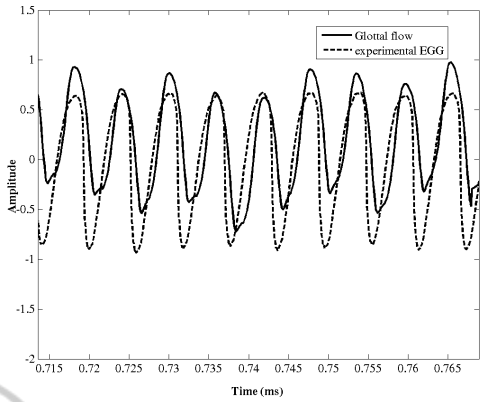

Figure 4: A matching between the EGG signal and the excitation source of a natural voice.

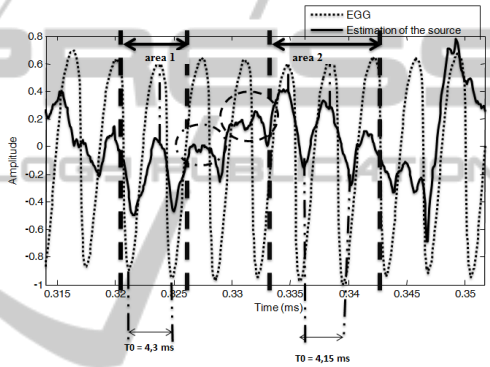

Figure 5: A matching between an artificial EGG signal (bold dashed line) and the excitation source (dashed line) of an esophageal voice.

Elsewhere, there are more difficulties to perform such identification of previous similarities. In fact, the fluctuations of the excitation source are so important that we can not identify the maxima of the excitation source (see for example circled areas in dashed ellipses in Fig. 5). We may call these cycles "deteriorated cycles". For example, from the instant 0.325 ms to the instant 0.33 ms, we can notice the absence of a clear maximum.

4 EXTRACTION OF RELEVANT INSTANTS IN THE EXCITATION SIGNAL

After using an artificial EGG signal to identify the cycles describing the behavior of the esophagus extremity, let's look for the significant instant of closing to better identify and characterize cycles within the esophageal speech signal.

In natural speech, the physical phenomenon of

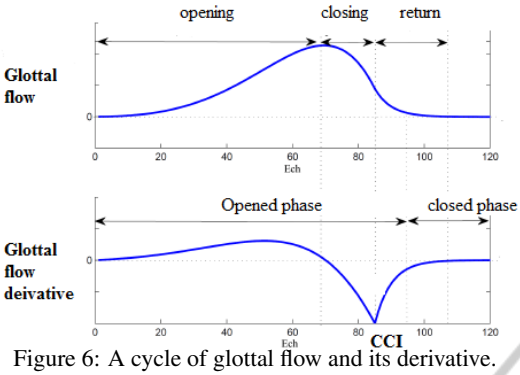

Figure 6: A cycle of glottal flow and its derivative.

voiced sounds pronunciation takes its origin in the lungs. Air passes later through the glottis and exerts a force on the vocal cords making them tick. The moment when the cords begin to separate is called Glottal Opening Instant (GOI). Similarly, after a maximum separation, the vocal cords tend to return to their original positions and the moment when the vocal cords re-stick is called Glottal Closure Instant (GCI). It corresponds to a minimum in the glottal flow derivative (see Fig. 6).

Is the same thing happens in esophagus extremity when trying to produce esophageal speech? This is the problem addressed now.

In order to identify the particular instant of closing GCI, we exploit the Dynamic Programming Projected Phase-Slope Algorithm (DYPSA) (Kounoudes et al., 2007). It is an automatic algorithm which operates using the speech signal alone without the need for an EGG signal. It incorporates a technique based on phase-slope function for estimating GCI candidates which are the instants of positive-going zero-crossings in the phase-slope function. Next, DYPSA algorithm employs dynamic programming to select the most likely candidates according to a defined cost function. This latter is a combination of five sub-functions considering pitch deviation, norm amplitude Consistency, ideal phase slope function, speech waveform similarity and projected candidate (Kounoudes et al., 2007).

When DYPSA algorithm is applied on natural speech, the GCI instants correspond perfectly to the instants when the excitation source derivative reaches the minima. In case of esophageal speech, we notice that the majority of the detected instants with DYPSA algorithm correspond to the minima of the excitation source derivative extracted with IAIF algorithm. But other candidates are badly placed in other positions and some candidates are missed (see Fig. 7 for example). This constatation can be explained by the fact there is not systematically complete opening and closing of the esophagus extremity. Consequently, there

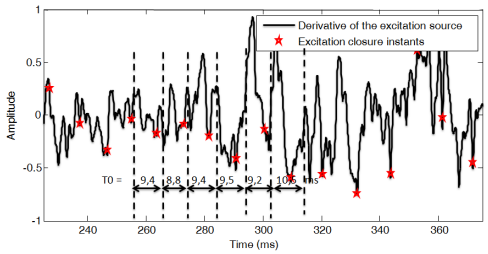

Figure 7: Superposition of GCI instants and the excitation source derivative.

Table 1: Pseudo-fundamental frequency for different phonemes of esophageal speech.

Phoneme 1 /a/	Phoneme 2 /a/	Phoneme 3 /o/
106 Hz	185 Hz	208 Hz
113 Hz	220 Hz	172 Hz
106 Hz	178 Hz	185 Hz
105 Hz	204 Hz	204 Hz
108 Hz	190 Hz	169 Hz

is no real cycles and real periods along the whole signal.

We'll give more details in next section.

5 IDENTIFICATION AND CHARACTERIZATION OF PSEUDO-CYCLES

In the central part of the signal of Fig. 7, some glottal cycles, defined as the period between two successive GCI, are detected. Their duration vary from 8.4 ms to 10.5 ms. They are called "pseudo-periods" as they change, even in a smooth way, from one cycle to an other.

In Tab. 1, we consider different phonemes, namely /a/ and /o/ and we attribute to some consecutive cycles a value of the fundamental frequency (inverse of the duration). This table shows that the fundamental frequency varies systematically from one cycle to its neighbor. It confirms the fact that the esophagus extremity vibration is not regular and not perfectly periodic, as it is the case of vocal folds.

In Fig. 8, we were able to match the identified cycles between two successive GCI to the cycles of an EGG generated at a frequency equivalent to the average of the pseudo-periods. The signal considered is the central part of the one in Fig. 7, whose beginning and ending are delimited by the the first and the last vertical dashed lines. The average frequency of 108 Hz is considered in Fig. 8 which corresponds to a pseudo-period of 9.2 ms. We recall that, according to

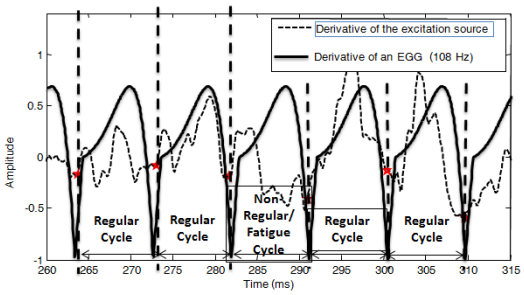

Figure 8: Classification of cycles according to esophagus extremity opening/closing.

Fig. 7, the real pseudo-periods range from 8.4 ms to 10.5 ms.

Moreover, it is important to note that, by considering this average fundamental frequency in Fig. 8, it is obvious that CGI instants does not correspond to the minimum of the excitation source derivative, since the fundamental real period is replaced by the averaged one.

After a deep study of an important number of esophageal speech sequences in terms of cycles structure and duration, one can go further in the analysis of cycles by defining a kind of cycle's classification. The proposed classification is illustrated in the signal of Fig.8.

- The two first cycles are considered as regular cycles or acceptable as they look more or less similar to the derivative of the EGG. Moreover, the local maximum of the cycle is identified as the maximum of the derivative of the EGG.
- The third cycle is classified as non-regular or fatigue cycle. This is a deteriorated cycle during which we cannot identify its maximum to the maximum of derivative of the EGG. It means that, the esophagus extremity was not able to open completely in easy manner and was trying to do it more longer in time than during regular cycles.
- The two following cycles behave more similar to the derivative of the EGG. It means that there is a return to regular cycles after a fatigue cycle.

In brief, we may explain getting a diversity of behavior by the fact that the esophageal voice, compared to the natural voice, is more difficult to produce and that the speaker is not fluent in producing this kind of voice. This may also explain the occurring of the so-called periods of fatigue during which the esophagus does not perform a correct cycle of opening/closing. This time interval of fatigue occurs after few proper cycles of opening/closing. After this time interval of fatigue which can be considered as a time interval of relaxing, the esophagus extremity recovers and produces more correct cycles. So, it has a behavior that could be a little recognizable or similar to the behavior of the vocal cords.

6 CONCLUSIONS

Even the complexity of the esophageal speech, the extraction of the excitation source permits to develop an analysis and a better comprehension of the esophagus extremity. We have, for example, shown some similarities between the EGG and the excitation source. Moreover, thanks to the localization of the CGI instants, we identified the equivalent of the glottal cycles in the esophageal speech. The classification of those cycles allowed to describe the status of the esophagus while producing the speech.

REFERENCES

Alku, P., Story, B., and Airas, M. (2004). Evaluation of an inverse filtering technique using physical modeling of voice production. In *Proc. of INTERSPEECH*. pp. 497-500, Kyoto-Japan.

Brown, D. H., Hilgers, F. J. M., Irish, J. C., and Balm, A. J. M. (2003). Postlaryngectomy voice rehabilitation: state of the art. In *World Journal of Surgery*. vol. 27, no.2, pp. 824-831.

Fant, G., Liljencrants, J., and Lin, Q. (1985). A four-parameter model of glottal flow. In *journal STL-QPSR*. vol. 4, pp. 1-13.

Kounoudes, A., Naylor, P. A., Gudnason, J., and Brookes, M. (2007). Estimation of glottal closure instants in voiced speech using the dypsa algorithm. In *IEEE Trans. Speech and Audio Processing*. pp. 34-43.