Esophageal Voices: Glottal Flow Restoration

B. García, J. Vicente, I. Ruiz, A. Alonso, E. Loyo

University of Deusto, Faculty of Engineering (ESIDE),
Av. Universidades 24, 48.007, Bilbao, SPAIN
e-mail: mbgarcia@eside.deusto.es, jvicente@eside.deusto.es, ibruiz@eside.deusto.es

ABSTRACT

One of the main problems when working with esophageal voices is the uselessness of traditional speech processing techniques. Concretely, when trying to delete noise from the speech signal, spectral subtraction techniques are not valid because the breathing noise which appears between syllables doesn't follow the same patterns as the one which appears during the speech. It is needed a new technique that allows to eliminate breathing noise, which seriously difficults later processing of this speech. The aim of this paper is to apply a new technique, which is based on a spectral match with esophageal noise patterns, to the treatment of noisy intervals between speech periods and the deleting of the glottal air flow characteristic spectral components.

1. INTRODUCTION

Among the different parameters accepted by scientific community to characterize the speech, as Pitch (fundamental frequency), Harmonics to Noise Ratio (HNR), Jitter or Shimmer, the second one is specially damaged [1] in esophageal speech. This effect is one of the most important ones and it is caused, among other reasons, by the involuntarily introduced air glottal flow [2]. That is the reason why it is produced the noise visible in the spectrum which requires a double treatment, which consists of the formants regeneration, as well as the breathing noise elimination in the no-sonority periods (it will be considered as sonority periods the frames with speech signal without considering if it is a voiced or an unvoiced frame). In this paper, this double processing is proposed.

The designed algorithm is able to detect the speech periods, while cleaning the non-speech intervals. Thanks to this, the obtained signal has its sonority intervals perfectly defined and ready for following treatment. Finally, the unpleasant effect produced by breathing noise in no-sonority periods is also deleted, achieving a clearer and more intelligible voice perception, which automatic voice recognizers can easily recognize.

2. OBJECTIVES

The main goal of the here presented algorithm is to enhance esophageal voice quality. Concretely, it will imply a higher HNR than the original one. As a consequence three particular purposes appear.

2.1. No-sonority frames detection

The first objective works on frames in the signal which are considered as no-sonority ones, that is, those frames which not contain information of the speech [3].

2.2. Breathing noise reduction algorithm design

This second objective consists on designing the algorithm which will be able to reduce the noise produced by the air going through the vocal tract without articulation.

An esophageal pre-processing algorithm will be achieved, so as the obtained signal is ready for the following processing.

2.3. Vocal tract model's poles stabilization

Finally, a notorious HNR improvement is achieved applying a designed algorithm for Vocal Tract model's poles stabilization over speech frames.

3. METHODS

The most important technique applied in the spectral analysis is the Fourier Transform, which will allow to recognize the spectral components of esophageal speech signal, so it makes possible to distinguish sonority intervals and process them. In order to achieve this, the signal will be analyzed spectrally frame by frame, so it will be possible to detect the frames on which the algorithm will be applied, which correspond with no-sonority ones, and the frames which are related to speech information. All of the algorithms are implemented in Matlab 6.5, as it provides the best framework for the design signal processing algorithm.

As it has been explained, this algorithm has a triple aim; in consequence, different types of speech processing techniques have been employed for each one.

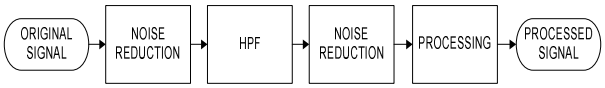

Fig. 1: High level flow chart.

On the one hand, the first objective involves the detection of no-sonority frames, in this sense, the use of pattern-recognition techniques makes it possible to detect those frames whose spectrums are rather similar to a reference noisy one, so as their effects can be attenuated. Due to the special characteristics of this kind of signals, it is impossible to apply traditional spectral subtraction methods to extract the breathing noise, because as it has been previously explained, its characteristics are rather different from the noise present in the sonority periods. That's the reason why a new type of noise detection algorithm has been designed.

On the other hand, with reference to the second objective, in order to reduce the breathing noise an algorithm to attenuate no-sonority frames has been developed.

Finally, in order to accomplish the third objective, the great variation of the poles' position in this kind of speech is corrected as shown in [4], achieving a HNR improvement.

4. DESIGN

The designed algorithm is carried out in two stages, as it can be seen in figure 1. The first one over the original esophageal voice signal, and the second one, after having applied a high pass filter to the output signal of the first step. This is so, because in the first iteration it's not possible to compare the noise reference frame with the low frequency one which appears all over the signal.

The noise reduction algorithm runs on several steps, as it can be seen in figure 2.

Initially, the original speech signal is normalized in time and pieced in several frames, this function runs on a subroutine which uses 1024 points windows, and hops of 512 points.

Over the several frames is carried out the following processing. The difference between its spectrum and the standard noise's one is calculated.

$$dif = \left| Gx(f) - Gn(f) \right|$$

$$dif\_mean = \frac{\sum_{n=1}^{m} dif}{m}$$

, where m is the number of points of the FFT, Gx(f) is the Power Spectral Density of the speech frame and Gn(f) is the PSD of the noise pattern.

If this difference is over a preestablished threshold, it can be concluded that both frames are different enough to consider the actual frame as a "not-noisy" one. In other case, the frame will be considered as noise. It is important to highlight that the process is no so simple, in order to improve the results of the algorithm, a special treatment is given to those frames which are classified as noisy ones, but are surrounded by speech periods. If this situation is detected the noisy frame will be considered as an exception and treated as a sonority one.

Having discriminated no-sonority frames, it is the aim of this work to attenuate them, in order to eliminate their effect. A study to calculate the correct factor of attenuation has been carried out, and as a conclusion of it, it has been determined that the worst sonority to no-sonority ratio is 3dB, as the minimum desired one to ensure speech quality is 10 dB (Noise magnitude is reduced to 10% of the speech signal), therefore the minimum attenuation factor for noisy frames should be at least 7 dB.

Finally, the reconstruction process is done and the resulted signal is normalized.

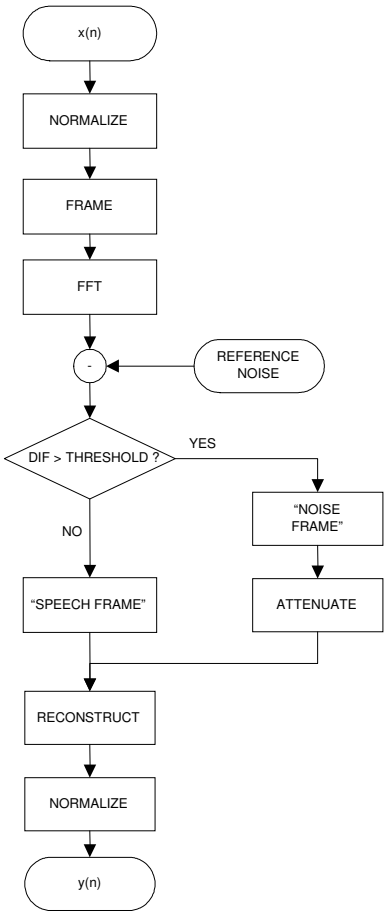

Fig.2: Low level flow chart.

A very important task consists on establishing an appropriated threshold so that the frame discrimination is correct. To achieve this, an empirical study with different speech signals has been developed, for this research, frames where an 80% of their length corresponded to speech signal and the other 20% could be considered as breathing noise, were used. The mean difference of these frames' spectrum with the noise pattern (dif_mean) has been fixed as the threshold.

In the following figure an esophageal speech signal is shown, the high pass filter has already been applied, so the low frequency noise is not present.

The differences between each sample of each frame's spectrum of the speech signal and the noise reference frame have been calculated and then the mean difference value determined. In the figure, sonority has been represented as a signal whose value is 'one' in speech frames, and 'zero' in those frames which can be considered as noise.

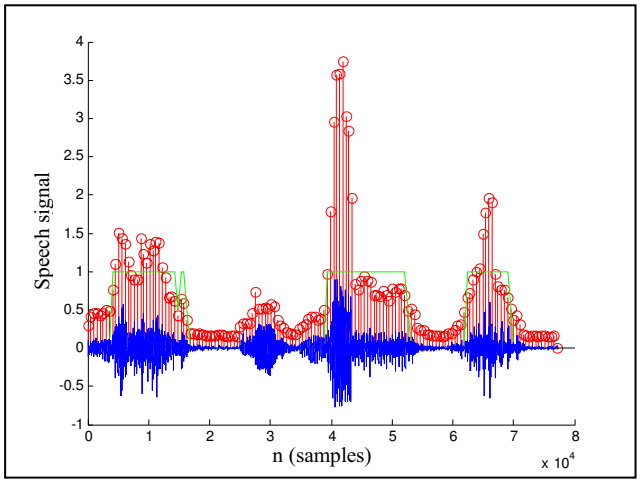

Fig. 3: Speech signal, differences means (circles) and sonority frames (continuous line).

Finally, a specific processing is applied over the "speech–considered" frames calculating the mean position of the three first formants and correcting their high variation.

5. RESULTS

As it can be observed in the following figure, thanks to the designed algorithm, a double objective has been achieved.

On the one hand, to eliminate low frequency noise, and on the other hand, the noise produced during the breathing, both of them typically appear in esophageal voices.

It has been mentioned that the "noise reduction" algorithm has to be applied twice to be effective because otherwise, some of the frames would not be correctly analysed. That is because of the low frequency noise, which has been removed using a high pass filter.

In step 2, after the second iteration of the algorithm, it is shown the cleaned signal. The breathing noise is completely removed and sonority frames have been detected correctly and processed, attending to the double objective it was pretended to achieve.

Finally, relating to the general purpose of the presented work, as a result of the implemented algorithm, a notable improvement of the Harmonics to Noise Ratio (HNR) of the signal has been reached, thanks to the attenuation in noisy frames and to the processing in speech ones. This can be traduced into a higher level of speech intelligibility.

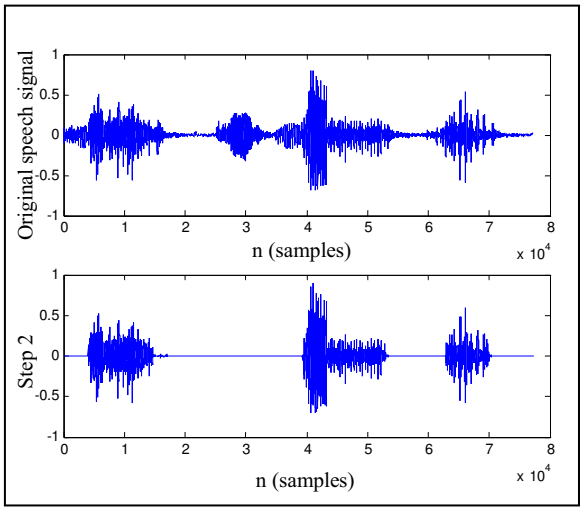

Fig. 4: Original vs. processed signal.

In order to demonstrate the improvement in this parameter, Figure 5 shows the relationship between the new value and the original value of the HNR. The diagonal line represents those cases where HNR maintains its original value after processing, in other words, the input signal without any processing. Thus, all values above the line will improve their HNR, as the value in the output signal will be higher than that in the input one, whilst those below the diagonal remain with a lower HNR.

This last case will correspond with the result of an incorrect application of the algorithm. This can be produced because of several causes, i.e.: an inappropriate discrimination threshold, a bad implementation of the high pass filter, or an incorrect attenuation factor.

According to the figure, there is no doubt that after applying correctly the performed algorithm, the HNR parameter improves in all cases, achieving the main objective of the work. The points over the line represent several esophageal speeches obtained from different laryngectomee speakers from the data base.

In all cases it is demonstrated that HNR reaches higher values after applying the algorithm.

In order to support this result, Table 1 reflects the improvement for 7 samples of esophageal speech from the data base. The obtained rank runs from 19% to 44%. These results prove the efficacy of the algorithm.

The mean enhancement for different voices is also shown in the following table, with a value of 28%.

As it has been shown, after the applying of the algorithm the speech signal is ready to apply over it any kind of additional speech processing in order to enhance its quality, as the silence signal periods have been perfectly delimitated, and it is possible now to work only over speech periods.

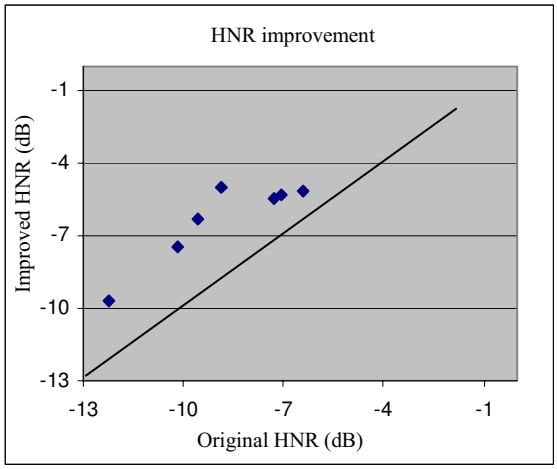

Fig. 5: HNR improvement for several voices.

6. CONCLUSSIONS

Despite of the several difficulties which appear when trying to work with these kind of speech pathologies, it has been demonstrated not only that it is possible to design an algorithm which achieves to eliminate the characteristic noise in silences periods of these signals, but also, that quality of esophageal voices can be improved.

It is very important to highlight that this algorithm does not remove completely the noise in speech frames, so it would constitute the aim of future investigations to design new algorithms for this task.

In this way, it is obtained a speech frame perfectly delimitated signal, so a later processing will be possible. It has to be mentioned also that some investigations in this line are already being carried out, trying to work on delimitated frames, result of the work developed in this paper.

It can be concluded that the enhancing work constitutes a great advance in the esophageal speech improvement process, in fact till now it had not been feasible to apply another similar algorithm to this kind of voices.

This is due to their special characteristics, and the fact that the breathing noise doesn't match with any other noise patterns, so commercial programs can not distinguish correctly 'sonority' or speech frames and 'no-sonority' or noisy ones. It was necessary to develop an algorithm according to their requirements and this duty has come across with several difficulties related with their extremely low pitch, high jitter and shimmer, and low HNR caused not only by the present noise, but also by the speech signal, characterized by weakness and low clearness.

Voices	Original HNR (dB)	Improved HNR (dB)	Improvement (%)
Voice 1	-12.252	-9.715	20.706
Voice 2	-7.079	-5.295	25.201
Voice 3	-10.148	-7.495	26.143
Voice 4	-6.398	-5.132	19.787
Voice 5	-9.578	-6.322	33.994
Voice 6	-8.877	-4.976	43.945
Voice 7	-7.312	-5.451	25.451
Mean			**28.296**

Table 1: Results after applying the algorithm.

Besides, this progress lets go on with the regeneration process working with words and even with full sentences. In other case, any attempt of enhancement using the already implemented algorithms, in stead of palliating the unpleasant effect produced during the breathing, would insert an important distortion in noisy periods. This fact would be damaging for the quality of the speech, worsening the HNR instead of improving it.

8. REFERENCES

[1] Baken, R. & Orlikoff, R. "Clinical measurement of speech and voice" Second Edition. San Diego, CA: Singular Publishing Group, ISBN:1565938690, 2000.

[2] Hooper, C.R. "Using evidence-based research in speech-language pathology: a project that changed my thinking" American-Speech-Language-Hearing-Association, January 2003.

[3] Brown, J.C. & Puckette, M.S. "A high resolution fundamental frequency determination based on phase changes of the Fourier transform" J. Acoust. Soc. Amer., vol.94, pp. 662-667, August 1993.

[4] García, B., Vicente, J. & Aramendi, E. "Time-Spectral Technique for Esophageal Speech Regeneration" Biosignal '02, 2002.