ANALYSIS AND TREATMENT OF ESOPHAGEAL SPEECH FOR THE ENHANCEMENT OF ITS COMPREHENSION

Jorge Miquélez, Rocío Sesma, and Yolanda Blanco

Department of Electrical and Electronic Engineering
ETSII & T, Public University of Navarra
Campus Arrosadía S/N, 31006 Pamplona, Navarra - Spain
E-mail: jomi@arrakis.es

ABSTRACT

This paper resumes an analysis of esophageal speech, and the developing of a method for improving its intelligibility through speech synthesis. Esophageal speech is characterized by low average frequency, while the formant patterns are found to be similar of those of normal speakers. The treatment is different for voiced and unvoiced frames of the signal. While the unvoiced frames are hold like in the original speech, the voiced frames are re-synthesized using linear prediction. Various models of vocal sources have been tested, and the results were better with a polynomial model. The fundamental frequency is raised up to normal values, keeping the intonation.

1. INTRODUCTION

Due to laryngeal cancer, and for its medical treatment, the removal of the larynx may be necessary. This will make the natural speech production mechanism to be different, since the air from the lungs cannot flow through the mouth and the nasal cavities. The windpipe must be brought out to the neck, to form a permanent opening called a stoma. Since the vocal cords are remove in the operation, or laryngectomy, the patient must learn a new manner of speaking.

After the operation, the vocal cords have been extirpated and the laryngectomized patients must use their esophagus to produce the speech. They pull into their esophagus a limited volume of air, which is thrown to the mouth for the speech production.

A linear model suggested in [1] has been chosen for the study, because it allows treating separately the different elements in the speech production (see fig. 1). A laryngectomy affects on the voiced sound generation mechanism, exactly, on their vocal source, modeled by the pitch period, the periodic excitation and the glottal pulse model. The operation does not affect on the vocal source of the unvoiced sounds and the formants of the phoneme, modeled by the random noise generator, and the vocal tract model, respectively.

2. ANALYSIS OF THE ESOPHAGEAL SPEECH

The fundamental frequency, or pitch, its fluctuations, the vocal source, and the vocal tract shape of two Spanish esophageal speakers have been analyzed and compared with those of a Spanish normal speaker. Isolated voiced sounds, isolated words, and short phrases have been recorded to analyzing and processing.

Esophageal speech is characterized by a very low average fundamental frequency. While the pitch is about 100-125 Hz for a normal male speaker, for a laryngectomized person of similar conditions is about 60 Hz.

In normal speakers, the pitch presents a soft variation through the discourse, because of the intonation of the sentence. The fundamental frequency of esophageal speakers shows a greater variability between consecutive periods of voiced frames, over a very short segment of time. This is due to the irregular vibration of the esophagus, and the hard control of this kind of vibration (see fig. 2).

The esophageal speech vocal source is not homogeneous in appearance. It has a like-noise, aperiodic shape, making the sound to be noisy and rough.

Otherwise, there are not important differences between the vocal tract of esophageal speakers and normal speakers. The study of the formants reveals that resonances of the tract appear at a no large range of frequencies for both laryngeal and alaryngeal speakers (see fig. 3).

[S0592_01.BMP]

Figure 1: Linear model of speech production.

3. TREATMENT OF THE ESOPHAGEAL SPEECH

The recordings were infected by a sinusoidal interference of 50 Hz, due to the Electrical Net. An adaptive filter was used to clean them, according to [2]. This was necessary because the fundamental frequency in esophageal speech is about 60 Hz, and the use of a fixed filter would have destroyed that information.

The recorded sentences were separated into voiced and unvoiced frames, with a standard deviation and zero-crossing based algorithm. The unvoiced frames were not treated. The vocal source in voiced frames was necessary to be replaced.

Over the voiced frames, a pitch-synchronous analysis was made to extract the Linear Predictive Coefficients of a 30^{th} order filter that models the vocal tract. The autocorrelation method, or Maximum Entropy Method (MEM), was used. This filter has the information about vocal tract.

The fundamental frequencies were calculated for each period of the voiced frame, and the obtained values were treated to make the synthetic transitions smoother, but keeping the intonation.

With the new smoothed fundamental frequency value, previously raised up to a normal one (around 100 or 125 Hz), one or two periods of vocal source were generated, and filtered with the LPC filter. Several waveforms were tested as vocal source, and a polynomial model introduced in [3] was chosen. This kind of excitation considers parameters like glottal open and close times, and its shape is similar to natural vocal source.

Both the unvoiced and voiced re-synthesized frames were finally connected to rebuild the original dialogue.

[S0592_02.bmp]

Figure 2: Fundamental frequencies for an esophageal speaker and a normal speaker, saying the Spanish word *martes*.

[S0592_03.BMP]

Figure 3: Central frequencies of the 3 first formants for the Spanish vowels, F1, F2, and F3. The values are similar for both alaryngeal and laryngeal speakers.

4. RESULTS AND EVALUATION OF THE METHOD

The results have been evaluated in two different ways: watching the spectrograms of the signals before and after the processing, and asking to people for the quality and the preference of both signals.

4.1. Spectral Characteristics of the Original and Synthesized Signals

The esophageal speech is characterized by a very low fundamental frequency average, and by a rough and noisy sound. Looking at the spectrogram of the recording signal *martes* (Tuesday, in Spanish), the formants can be appreciated clearly. But the typical spectral lines due to the pitch, and its harmonics do not appear (see fig. 4.a).

After the processing, the periodicity due to the pitch, and its harmonics, appears in the synthesized signal. This is shown in the spectrogram by horizontal spectral lines in the voiced intervals (see fig.4.b). In consequence, the sound is clearer and the speech is more intelligible than the original one.

4.2. Acoustic Evaluation

Seven Spanish words pronounced by a laryngectomized patient were taken and processed according to the method here proposed. The words were *lunes*[S0592_01.WAV][S0592_08.WAV], *martes*[S0592_02.WAV][S0592_09.WAV], *miércoles*[S0592_03.WAV][S0592_10.WAV], *jueves*[S0592_04.WAV][S0592_11.WAV], *viernes*[S0592_05.WAV][S0592_12.WAV], *sábado*[S0592_06.WAV][S0592_13.WAV], and *domingo*[S0592_07.WAV][S0592_14.WAV].

Twenty undergraduate students at the Public University of Navarra were asked to listen to the synthesized signals. Without any previous knowledge about the signals, they had to answer four questions:

1. Which one of the signals seems to be rougher?
2. Which one has a better intonation?
3. Which one is thought to be the original recording?
4. Which one is preferred to listen?

[S0592_04.BMP]

[S0592_15.WAV]

[S0592_16.WAV]

Figure 4: Spectrograms of the word *martes* a) said by an esophageal speaker, and b) re-synthesized according to the method proposed here.

The result of the quiz was:

1. 95.00% of the people queried thinks the synthesized signal is less rough than the original one.
2. 72.14% thinks the synthesized signal intonation is the better one.
3. 93.57% can recognize the esophageal speech.
4. 70.00% prefers to listen to the synthesized signal.

If an average of these data is made, the method is considered as good, with a ration of 82.68%.

5. CONCLUSION

This paper propones a method to improve the acoustical quality of esophageal speech. The method works over words and sentences. Automatically, it sorts the frames into voiced and unvoiced ones, and treats them to obtain a synthesized signal, with better characteristics.

The results are very satisfactory, and hopeful. A real-time implementation of the method could be a very useful tool for esophageal speakers.

A bigger data collection to study would be desirable to obtain a more standard conclusions and results.

6. ACKNOWLEDGMENT

The authors of this paper are so grateful to Dr. Javier Medina, for his support and interest, and to Javier Sánchez, Isabel Aranguren and Hilario Iza, for their availability and happiness.

7. REFERENCES

1. Oppenheim, A., and Schafer, R. *Discrete-Time Signal Processing*, 815–821, Prentice–Hall, 1989
2. Widrow, B., and Stearns, S. *Adaptive Signal Processing*, 302–367, Prentice–Hall, 1985.
3. Van Santen, J., Sproat, R., Olive, J., and Hirschberg, J., *Progress in Speech Synthesis*, 27–39, Springer, 1996.