Audio Engineering Society
Convention Paper
Presented at the 121st Convention
2006 October 5–8 San Francisco, CA, USA

Esophageal Voice Enhancement by Modeling Radiated Pulses in Frequency Domain

Alex Loscos[1], Jordi Bonada[1]

[1] Music Technology Group, Universitat Pompeu Fabra, Barcelona 08003, Spain
alex.loscos@iua.upf.edu

ABSTRACT

Altough esophageal speech has demonstrated to be the most popular voice recovering method after laryngectomy surgery, it is difficult to master and shows a poor degree of intelligibility. This article proposes a new method for esophageal voice enhancement using speech digital signal processing techniques based on modeling radiated voice pulses in frequency domain. The analysis-transformation-synthesis technique creates a non-pathological spectrum for those utterances featured as voiced and filters those unvoiced. Healthy spectrum generation implies transforming the original timbre, modeling harmonic phase coupling from the spectral shape envelope, and deriving pitch from frame energy analysis. Resynthesized speech aims to improve intelligibility, minimize artificial artifacts, and acquire resemblance to patient's pre-surgery original voice.

1. INTRODUCTION

Most laryngectomized patients suffer from voice blackout after neck surgery, having mainly two ways to recover voice. One consists in using an Artificial Larynx, a hardware device which held against the neck or cheek produces sound vibrations in the throat, while the speaker articulates with the tongue, palate, throat and lips as he does for the usual vocalization. Those devices can be easily mastered, but the sound quality is rather electronic and artificial. Besides, one hand is employed to hold the device during speech, disturbing the gestural communication.

The alternative way consists in training esophageal speech, a way of speech production based on the technique in which the patient transports a small amount of air into the esophagus. Probably due to an increased thoracic pressure, the air is forced back past the pharyngo-esophageal segment to induce resonance and allow speech. Rapid repetition of such air transport can produce understandable speech. However, on average, esophageal voice results in low-pitched (~50Hz), low intensity speech, and with a poor degree of intelligibility. On the other side, this technique is able to preserve the speaker's individuality, since the speech is generated by his own vocal organs, and the speaker is

able to use his hands and facial expression freely and actively for a more natural communication.

Several researches which pretend to enhance and clarify the esophageal speech have been reported so far [1][2], as well as studies which account for its lack of phonetic clarity [3][4]. Besides, a device with the aim of improving the esophageal voice is commercially available since few years ago [5] This device consists of a small circuit board, a compact microphone and a speaker, and uses formant synthesis techniques to produce the synthetic voice. One of its main disadvantages is that it fails to keep the speaker's individuality.

We propose a straightforward digital signal processing system based on analysis, transformation and resynthesis for esophageal voice enhancement. The new algorithms we describe in this article have been designed to:

a) improve intelligibly of esophageal speech

b) allow laryngectomized patients have intuitive control on the resynthesized prosody

c) minimize traces of artificialness

d) and resemble pre-surgery patient healthy voice.

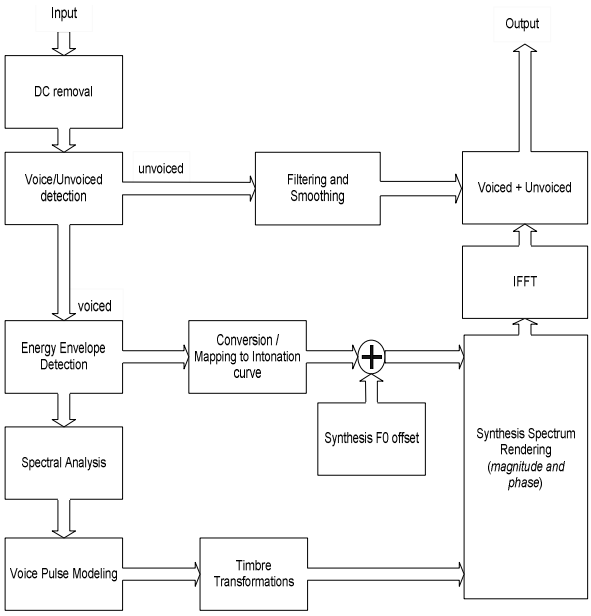

Figure 1 General block diagram of the Voice Enhancer

2. ESOPHAGEAL VOICE ENHANCER

The proposed voice enhancer is based on the Voice Pulse Modeling (VPM) algorithm [6] which intends to combine the waveform preservation ability of time-domain techniques with the flexible and wide transformations of frequency-domain techniques, while at the same time avoiding the complexity and the contextual problems of them.

Our implementation of the system is based on frame analysis. After filtering, the input voice frame classified as voiced or unvoiced. For the voiced frames spectral phase and magnitude (timbre and pitch) information is made up to feed the VPM. The voiced spectrum is then converted to time domain by the IFFT module and added to the processed unvoiced utterances.

3. THE PREPROCESS AND THE UNVOICED CYCLE

The preprocess block at this point of research consists on a simple high pass filter (DC removal in figure 1). The filter is applied to get rid of the very low frequencies (from 10 Hz to 30 Hz approximately) typically present in esophageal voice recordings due to powerful respiration.

Once filtered, there is a voiced / unvoiced detection which does not rely on fundamental frequency analysis as it does traditionally, but it bases on a combination of signal dynamics and spectral centroid. For those frames that are considered unvoiced, second step equalization is applied in order to reduce unvoiced consonant perceptual forcefulness.

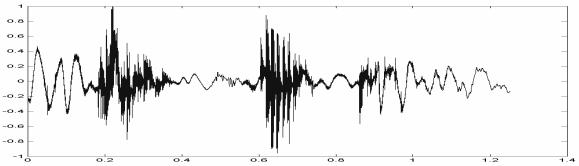

Figure 2 Waveform (normalized magnitude versus sample index) of the recording of the Spanish word 'tomate' uttered by a laryngectomized patient

4. VOICE PULSE MODELLING

If voiced, the utterance is analyzed and resynthesized with new timbre, phase and prosody. Because of the rough nature of esophageal speech, no pitch analysis is

performed. Thus, spectral timbre envelope and frequency phase alignment are not computed using harmonic peak information as usually done in the VPM technique (figure 3).

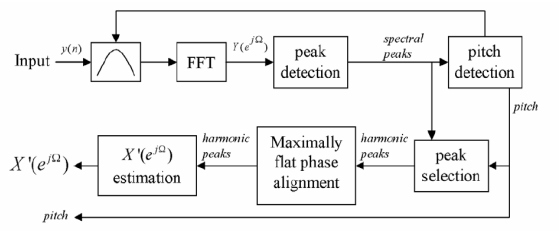

Figure 3 Block diagram of the VPM analysis process taken from [6].

For the case of the timbre, we track spectral envelope using a bank of non-overlapped constant bandwidth (~175 Hz) filters equally spaced in frequency. Frequency phase alignment envelope is derived instead directly from the resulting dB magnitude envelope by low pass filtering, offsetting, shifting and scaling its y-axis to fulfill phase 0 at frequency 0 and approximately pi phase drops under most prominent formants. A smoothing function ensures no phase alignment discontinues at consecutive frames.

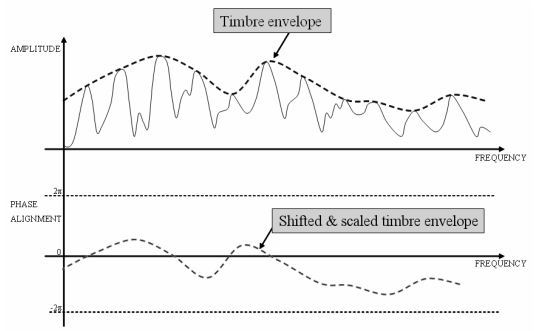

Figure 4 Hand drawn representation of phase alignment envelope generation

These frequency magnitude and phase envelopes are then used by the VPM to model the frequency response of a single voice pulse centered in the analysis window. Next, together with the pitch the rendering module generates the synthesis frequency domain spectrum out of the transformed voice pulse models.

Synthesis pitch envelope $f_0(t)$ is obtained by means of filtering, scaling and offsetting the energy envelope (in dB scale) detected in the analysis:

$$f_0(t) = filter\left[10 \cdot \log\left(s^2(t)\right)\right] \cdot a + b \qquad (1)$$

Where a=12.5 and ß=-2400 have been found to be appropriate values for the male voice used as an example. However, values should be tuned for each specific case in order to fit as much as possible original patient's average speech fundamental frequency and prosody intonation.

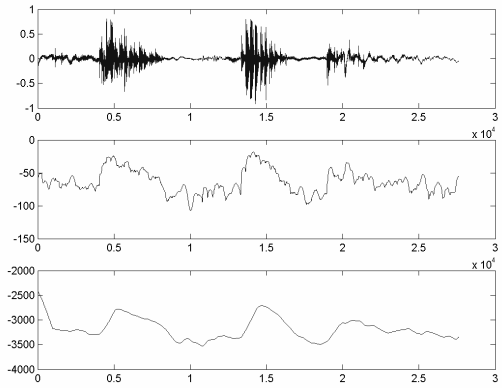

Figure 5 Along sample index, upper plot shows the original input signal after preprocess, mid plot shows its energy envelope in dBs, and lower plot shows the computed healthy synthesis pitch in cents of a tone.

5. HEALTHY PITCH AND TIMBRE

Although alaryngeal speakers suffer mainly from changes in those characteristics related to the voice source, relevant changes occur as well in the vocal cavity transmission characteristics.

Recent studies in which acoustical analysis of vowels was carried out show alaryngeal speakers place their formants in higher frequencies. According to [7] the explanation of this symptom seems to be that total laryngectomy results in shortened vocal tract relative to normal subjects. In fact, Diedrich and Youngstrom [8] demonstrated using data of a patient captured before and after laryngectomy that effective vocal tract length is reduced after neck surgery.

In order to compensate the aforementioned formant shift, the resulting bank magnitude envelope is frequency-varying scaled using a timbre mapping function such as:

$$\left|X_s(f)\right| = \left|X_a\left(f^a\right)\right| \qquad (2)$$

where $|X_s(f)|$ is the synthesis spectral envelope and $|X_a(f)|$ the spectral envelope obtained in the analysis.

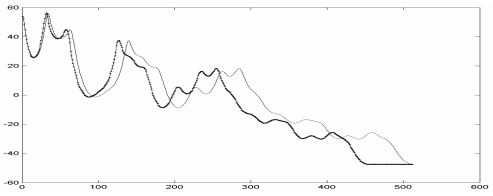

Figure 6 Log spectral magnitude versus frequency index plot of analysis (lighter) and synthesis (darker) timbre envelopes for an [o] utterance

For values of a around 1.02 the frequency down shift achieved is around 60 Hz for the first formant and 160 Hz for the second formant, which are the values spotted in [7] as average formant rise values.

6. CONCLUSIONS

Results prove the proposed automatic technology is a first step towards an esophageal voice enhancer based on digital signal analysis, transformation and resynthesis, able to resemble pre-surgery voice; an enhancer that laryngectomized patients could get to feel as being part of their voice production mechanism instead of seeing it as an outside effector.

The system has been specifically designed to work in real time. Although probably far from it, we aim for real time esophageal voice enhancers to be integrated into small portable devices. Such tools would definitely improve laryngectomized patient's quality of life.

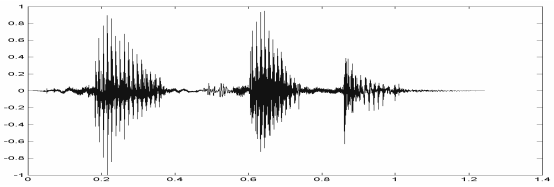

Figure 7 Waveform (normalized magnitude versus sample index) of the resynthesized 'tomate' utterance represented in figure 2

7. ACKNOWLEDGEMENTS

We would like to acknowledge the help and support received from "Asociación Barcelonesa de Laringectomizados", "Centro de Rehabilitacion para Laringectomizados de Terrassa", "Escuela de Patologia y Lenguaje del Hospital de la Santa Cruz y San Pau", "Pla Director d'Oncologia de Catalunya", and Begoña García from University of Deusto.

This work was partially supported by the European Union's 6th Framework Programme's project SALERO.

8. REFERENCES

[1] Sawada, H, Takeuchi, N, Hisada, A, 'A Real-time Clarification Filter of a Dysphonic Speech and Its Evaluation by Listening Experiments, International', Conference on Disability, Virtual Reality and Associated Technologies (ICDVRAT2004), pp. 239-246, 2004

[2] Doi, T, Nakamura, S, Lu, J, Shikano, K, 'Improvement in oesophageal speech by replacing excitation components in cepstrum domain', The Acoustical Society of Japan, Autumn Meeting 2-4-17, pp. 253-254, 1996

[3] Bellandese, M, Lerman, J, Gilbert, J, 'An Acoustic Analysis of Excellent Female Oesophageal, Tracheoesophageal and Laryngeal Speakers', Journal of Speech, Language and Hearing Research, 44, pp. 1315-1320, 2001

[4] Robbins, J, Fisher, H, Blom, E, Singer, M, 'A comparative acoustic study of normal, oesophageal and tracheoesophageal speech production', Journal of Speech and Hearing Disorders, 49, pp. 202-210, 1984

[5] '2005 Romet Electronic Larynx product borchure' http://www.larynxlink.com/suppliers/RometBrochure.pdf

[6] Bonada, J, 'High Quality Voice Transformations Based On Modeling Radiated Voice Pulses In Frequency Domain', Proceedings of 7th International Conference on Digital Audio Effects Naples, Italy, 2004

[7] Cervera, T, Miralles, J.L, González, J, 'Acoustical Analysis of Spanish Vowels Produced by Laringectomized Subjects', Journal of Speech, Language, and Hearing Research, 44, 988-996, 2001

[8] Diedrich, W.M., Youngstrom, K. A., 'Alaryngeal Speech', Springfield, IL: Charles C. Thomas, 1966.