Available online at www.sciencedirect.com
ScienceDirect
Biomedical Signal Processing and Control 1 (2006) 107–112
Biomedical Signal Processing and Control
www.elsevier.com/locate/bspc

Methods for formant extraction in speech of patients after total laryngectomy

Rafał Pietruch [a,*], Marta Michalska [b], Wiesław Konopka [b], Antoni Grzanka [c]

[a] *Medical University of Warsaw, Department of Prevention of Environmental Hazards, ul. Banacha 1a, 02-091 Warsaw, Poland*
[b] *Medical University of Łódź, Department of Otolaryngology, ul. Zeromskiego 113, 90-549 Łódź, Poland*
[c] *Warsaw University of Technology, Institute of Electronic Systems, ul. Nowowiejska 15/19, 00-665 Warsaw, Poland*

Received 17 February 2006; received in revised form 22 September 2006; accepted 22 September 2006

Abstract

The paper shows the methods and its application for voice analysis suited to the group of subjects after total laryngectomy surgery. Our software was developed to evaluate and enhance laryngectomized patients' rehabilitation process. The power spectral density imaging and formant frequencies extraction methods were adopted. The model of vocal tract was based on statistical, autoregressive process of speech production. The transversal filter and adaptive algorithm were implemented to estimate the transfer function of resonance cavities. The research is concerned with measurements of vowel articulation parameters, especially F1 and F2 formant frequencies. The significant difference of pathological and normal voice in vowel space separation has been presented. The authors found that formants in pseudowhisper speech are more pronounced while articulating vowel after consonant than for sustained vowel.

Keywords: Speech analysis; Autoregressive model; RLS algorithm; Formant extraction; Total laryngectomy; Pseudowhisper

1. Introduction

Laryngectomy is a partial or complete surgical removal of the larynx, usually performed as a treatment for laryngeal carcinoma. Cancer of larynx is the most common head and neck tumor in males (17 times more cases than in women) and one of the most common malignancies in Europe. This tobacco dependent neoplasm accounts for more than 5% of all cancers among men in the average age of almost 60 years [1]. However, compared to other malignancies, the prognosis is favorable at 60% survival rate. Surgical removal of the larynx remains the procedure of choice for advanced stage laryngeal carcinoma T3 and T4 Union Internationale Contre le Cancer (UICC) [1,2]. However, one must keep in mind the drawbacks of any surgical procedure. In this case these may include loss of nasal function [3], changes of lung function, poor cough, swallowing difficulties and tracheostomal complications [2]. As a consequence there is also a major impact on patient's psychology and his life.

Following the loss of vocal cords patients are not able to phonate adequately. The biggest problem for laryngectomees is to pronounce vocalized sounds that are naturally articulated with the use of vocal cords. Their voice is hoarse, weak and strained [4], therefore the main goal of phoniatric rehabilitation is to teach patients how to articulate understandable speech. During the therapy subjects learn how to force pharyngo-esophageal segment to articulate alternative sounds that substitute fundamental frequency [2]. A certain percentage of laryngectomees never acquires an alaryngeal voice neither is able to use an electronic larynx. They usually communicate using pseudowhisper, by silently articulated words with some ejectives from intra-oral pressure [5].

It is well known that vowels are identified mostly through their formant frequencies [6], and therefore a major part of the perceptual information contained in voiced speech is encoded in those values [7]. The algorithm proposed in this paper attempts formants extraction from voice signals. The authors concentrated on the differences between the formant frequencies of normal and pathological voice. Differences between alaryngeal and natural speech have been already studied [8–10]. We wanted to compare the pseudowhisper with natural and esophageal voice to find the disorder characteristics.

* Corresponding author. Tel.: +48 22 599 21 86.
E-mail address: rpietruch@am.edu.pl (R. Pietruch).

1746-8094/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bspc.2006.09.001

The paper presents voice analysis methods and studies their application in a group of subjects after total laryngectomy. An adaptive algorithm presented in Ref. [11] has been implemented to estimate the vocal tract model parameters. The power spectral density imaging and extraction of formant frequencies methodology was adopted to compare natural voice with esophageal and pseudowhisper speech. A personal computer software was created for real time visualization of the time-varied PSD of the speech and formants tracking. In order to improve and simplify medical analysis the research attempt to provide an objective measure of postoperative voice quality and phoniatric rehabilitation enhancement. Statistical analysis was used to evaluate patients' voice formant separation.

2. Material and methods

Fifty Polish-speaking patients who had undergone the total laryngectomy and 10 Polish-speaking normal volunteers forming the control group participated in this project. Clinical and demographic distribution is shown in Table 1. To facilitate statistical analysis the patients were grouped according to the time after surgery, and whether or not they underwent radiotherapy and speech rehabilitation. All patients used an esophageal speech or articulate with the use of silent mouthing ejectives and pseudowhisper voice.

Recordings were made in a sound-proof booth using a digital video camera (Panasonic NV-DS65EGE) and an external, battery operated electret-condenser microphone with custom gain adjustment circuit. The microphone was mounted in front of the speaker's mouth at a distance ranging from 15 to 20 cm and at an angle of about 45°. Subjects where asked to read twice the linguistic material written on cards. The utterances comprised 6 Polish vowels, 10 words and 2 sentences (see Appendix A) designed according to the guidelines of Zakrzewski et al. [12]. All words were in Polish in the format of consonant–vowel–consonant (CVC) or consonant–vowel–consonant–vowel (CVCV).

The material was recorded on MiniDV tape. The Pinnacle Studio video card was used to transfer the recordings to a computer. Audio signals were down-sampled from 44.1 kHz down to 8 kHz sampling rate using CoolEdit Pro 2.0 software and stored on a computer disk in WAV format. The signal-to-noise ratio (SNR) was 45 dB as calculated by MATLAB software. The SNR was calculated using 0.25 s long fragment of loud speech and a 0.25 s long recording of 'silence'.

The authors developed software that visualizes PSD of speech and helps to track the formants from audio WAV files. Statistical auto-regressive (AR) modeling approach was chosen for speech dynamics analysis with an assumption that human speech is a linear transformation of white noise. Specifically, a digital infinite impulse response (IIR) filter equivalent to natural vocal tract transmittance [13,11] was used. The filter transmittance $H(z)$ is described by the following Eq. (1) [7]:

$$\hat{A}(n, f) = |\hat{H}(n, f)| = \left| \frac{\hat{Q}(n)}{1 - \sum_{k=1}^{p} \hat{h}_k(n) z^{-k}} \right| \quad (1)$$

where f_s stands for frequency, n a sample number and z is defined as $z = \exp(2\pi i f/f_s)$. Note that transmittance H is a function of time, according to equation $n = tf_s$. Variable $Q(n)$ is a temporal power of windowed prediction error that represents the power of acoustical wave sources signal [11].

Linear prediction LP has been used for estimation of the inversed transversal filter parameters. The LP coefficients have equivalents in the AR parameters h_k. The spectrum of speech signal was calculated from the vocal tract filter coefficients. The number p of estimated parameters of the filter was set to 8 for vowels analysis from signals sampled with 8 kHz, for pseudowhisper speech the filter order was set to 10. In order to check if formants were not blended (did not split into two) the number of filter coefficients was increased above 10. This method performed particularly well with Polish vowels 'o' and 'u' where F1 and F2 formants are very close and can be fused for low model order. Calculation of the estimated filter parameters h_k for nth sample was performed using weighted recursive algorithm based on least-squares error minimization (WRLS) [11]. The weighting parameter λ in the algorithm was matched experimentally by authors and set to 0.985.

Calculation of the covariance matrix and LPC was performed every sample. The AR polynomial was updated every 50 samples (about 6.25 ms for $f_s = 8$ kHz), and the spectrum visualized as a new vertical colored line. Frequency resolution of the chart was described by the formula $f_s/640$ (12.5 Hz for 8 kHz). In order to equalize the overall energy distribution the natural and esophageal speech signals were pre-emphasized using a high-pass finite impulse response (FIR)

Table 1
List of patients' clinical data

Variable	Pseudowhisper ($n = 33$)	Esophagel speech ($n = 17$)	Control group ($n = 10$)
Sex, no. (%)			
M	22 (67)	16 (94)	5 (50)
F	11 (33)	1 (6)	5 (50)
Age (years)			
Mean (S.D.)	62 (9)	64 (10)	
Median	61	63	
Range	44–83	49–79	25–50
Radiotherapy, no. (%)	19 (58)	8 (47)	
Speech therapy, no. (%)	19 (58)	11 (65)	
Time after surgery, <5 years, no. (%)	26 (79)	8 (47)	