METHODS FOR FORMANT EXTRACTION IN SPEECH OF PATIENTS AFTER TOTAL LARYNGECTOMY

R. Pietruch[1], M. Michalska[2], W. Konopka[2] A. Grzanka[1,3]
[1]Institute of Electronic Systems, Warsaw University of Technology, Warsaw, Poland
[2]Department of Otolaryngology, Medical University of Lodz, Lodz, Poland
[3]Department of Prevention of Environmental Hazards, Medical Academy of Warsaw, Poland

Abstract: **The paper shows the methods for imaging power spectral density of speech and extracting formant frequencies from continuous voice. The methods will be used to improve the patients' rehabilitation after the total laryngectomy surgery. The adaptive algorithms and transversal filters were implemented to estimate the transfer function of human vocal tract model. The estimation methods were based on statistical, Auto-Regressive model of speech production. The pilot study on formant frequencies, especially F1 and F2 formants, and their linear separation for each vowel has been presented. The method for recognition pathological voice has been proposed.**

Key words: **Speech signal spectrum analysis, adaptive filters, formant tracking and total laryngectomy**

I. INTRODUCTION

Laryngectomy is a partial or complete surgical removal of the larynx, usually performed as a treatment for laryngeal cancer. After the loss of vocal cords patients are not able to vocalize their speech. It is difficult for them to generate phonation, which would be understandable and communicative. Their voice is hoarse, weak, and strained. The main goal of phoniatric rehabilitation is to teach patients how to articulate understandable speech. During the therapy subjects are learning how to force pharyngo-esophageal segment to induce resonance and articulate alternative voice. Esophagus should become a vicarious source and substitute vocal cords. A certain percentage of laryngectomees never acquires an alaryngeal voice and is unable to use an electronic larynx. They usually communicate by silently articulated words with some ejectives from intra-oral pressure. This voice called silent mouthing is not a truly oesophageal voice.

To improve and simplify the medical analysis the computer program has been written. It visualizes power spectral density (PSD) of speech. The algorithm presented in this paper has been implemented to estimate PSD and to track formants from continuous speech in real time. The estimation of vocal tract model parameters and formants extraction was used for comparing natural voice with oesophageal and silent mouthing speech.

II. METHODOLOGY

A. Linear prediction and adaptive filters

The algorithm proposed in this paper attempts formant extraction from voice signals. The tracking formant algorithms have been proposed e.g. in papers [1], [2]. In presented applications the Auto-Regressive (AR) statistical process models speech dynamics. It was assumed that human speech is a linear transformation of white noise. AR process was used instead of Auto-Regressive Moving Average (ARMA) model. Thus, the voice spectrum analysis was simplified and nasal tract transmittance influence was eliminated. The digital Infinite Impulse Response (IIR) filter equivalent to natural vocal tract transmittance [3] models the AR process. The filter transmittance H estimated for n-th sample, depends on signal frequency f, it is a function of complex number $z = e^{2\pi \cdot j \frac{f}{f_S}}$, where f_S is a sampling frequency. According to [4, 5], the amplitude of H is given by equation (1):

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

H transmittance can be presented in time units, according to equation $n = tf_S$. Variable Q(n) is a temporary power of prediction error, which is analogical to the power of signal generated with larynx or noise produced by air turbulences [6].

The linear prediction has been applied to estimate the inversed transversal filter parameters. Linear prediction coefficients (LPC) are the AR parameters h_k. The transversal filter was used instead of lattice filter because of simpler numerical complexity. However the PARCOR reflection coefficients can be computed from the LPC [7].

Models and analysis of vocal emissions for biomedical applications. 4th international workshop. October 29-31, 2005 – Firenze, Italy.
Edited by C. Manfredi. ISBN 88-8453-320-1 (online) © 2005 Firenze University Press

B. Recording Procedures

Fifty Polish-speaking patients who had undergone the total laryngectomy and twenty subjects from control group were recorded with the use of a digital camera, Panasonic NV-DS65EGE. The recordings were made in the sound-treated booth in order to minimize the background noises. An electret-condenser microphone was connected to the camera and supplied with the R6, 1.5V battery. A potentiometer between microphone and camera input was used to adapt the signal power. The microphone was positioned on a clip mounted around a neck. The distance between mouth and microphone was about 15cm. The linguistic material was presented on cards. Sentences were read twice.

The audio-video material was recorded on MiniDV tape. The Pinnacle Studio video card was used to transfer the recordings to computer for acoustic analyses. The data was stored on MPEG format files. The audio signals were digitized with 44,1 kHz sample rate. The Signal-to-Noise ratio (SNR) of recorded tracks was about 45dB, according to calculations on MatLab 6.0 application. The SNR was calculated as the proportion of the power of silent and speech signals based on 1000 selected samples. The actual range of speech signal power is about 60dB [8].

A computer program was written to visualize spectral density and track the formants of human speech. It processes the audio data from WAVE and MPEG multimedia files. For WAVE format files audio signals were decimated down to 8 kHz sample frequency using CoolEdit Pro 2.0 software. For the vowels analysis based on two first formats the 8kHz sample frequency was used because these formats appear in the 4kHz range.

C. Spectrum estimation and tracking formants

The spectrum of speech signal is calculated from the vocal tract filter coefficients. The number of estimating filter parameters h_k, can be changed in the program according to the sampling frequency f_s and the nature of voice. For vowel analysis from signals sampled with f_s = 8kHz the number of LPC was set to k = 8. According to the literature [3, 8], up to four formants should be placed in the 0-4kHz-frequency range. To check if formants are not blended the number of filter coefficients was increased to k = 16. It was checked then if one formant didn't split into two. This method had significant results especially for Polish vowels /o/ and /u/ where F1 and F2 formants are very closed.

Calculation of the filter estimated LPC parameters h_k in n-th sample is based on Recursive algorithm based on Least-Squares error minimization (RLS). Haykin listed detailed steps of algorithm in [9]. The constants in algorithm were matched experimentally by authors and set as: $\alpha = 0.05$, $\lambda = 0.985$.

By experimental research it is proven that RLS algorithm is characterized by a fast rate of convergence. According to the literature [9] the mathematical formulation and therefore the implementation of RLS is relatively simple and efficient in computation. In [9] Haykin shows a numerical instability problem considered in finite precision arithmetic. Applied method cures the divergence of standard RLS algorithm.

According to experiments, listed algorithm is suitable for a real time implementation on the personal computer for sound data sampled with up to 44kHz frequency.

The amplitude of speech spectrum has an exponenttially falling character, about -12dB per octave [3]. To equalize the overall energy distribution the speech is pre-emphasized using a high-pass FIR filter with parameter vector h = [1 -0.9735].

According to Christensen method [10] the local minimums of second derivative of power spectrum A(m,n) were searched for the formants extraction. This method can separate some blended formants. It was assumed in algorithm that the second derivate must be negative.

III. RESULTS

A. Tracking formants results.

The time-frequency spectrograms are presented in Figures 2, 3 and 4. The level of gray is proportional to the amplitude. Local minimums were colored black for better vision.

The comparison of pathological (Fig.2, 3) and natural voice (Fig.1) shows the huge differences in the articulation of speech. Shorter articulation of vowels, and higher noise level can be seen for esophageal voice (Fig.2). The pauses are longer. However the formants of esophageal voice match (with little variation) the natural speech spectrum (Fig.1).

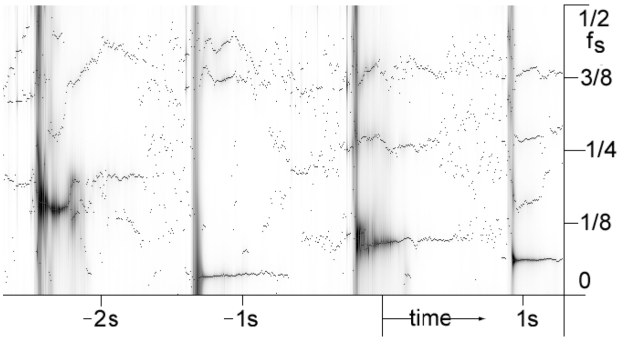

Fig.1 Spectrogram of natural voice (Polish vowels /a/, /i/, /e/ /y/), X axis: time, Y axis: f/f_S, f_S= 8kHz.

For silent mouthing (Fig.3) the differences between each vowel formants are insignificant Thus, it is hard to recognize each vowel from the spectrum. Presence of noise is more evident than in the alaryngeal group.

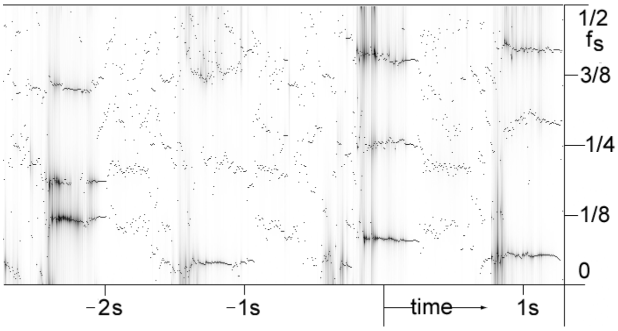

Fig.2 Spectrogram of esophageal speech (Polish vowels /a/, /i/, /e/ /y/), X axis: time, Y axis: f/f_S, f_S= 8kHz.

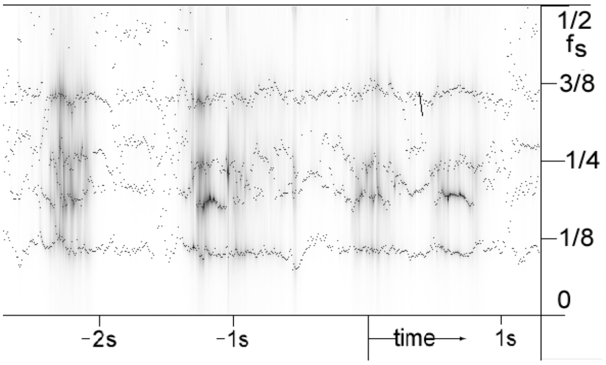

Fig.3 Spectrogram of silent mouthing voice (Polish vowels /a/, /i/, /e/ /y/), X axis: time, Y axis: f/f_S, f_S= 8kHz.

It is evident that there are higher frequencies of first formant for silent mouthing vowels articulation (Fig. 3). In this speech no fundamental frequency excitation sources are involved in speech production. Although the vocal tract parameters are similar, the transfer function is different because of other source of air turbulences [3]. Moreover, the speech is distorted and the spectrum covered by the air turbulences noise from tracheotomy tube and its spectrum with regular resonances. Thus, the spectrum is distorted and the speech less intelligible.

B. Vowels classification.

It is well known that vowels are identified mostly through their formant frequencies [9] and therefore a major part of the perceptual information contained in voiced speech is encoded in these formant frequencies [7]. This paper is concerned with the differences between the formant frequencies of normal and pathological voice.

Mean values and deviations of the first and second formants of the vowels produced by normal, esophageal and silent mouthing voices have been presented in Figures 4, 5 and 6. The measurements were performed for 3 subjects of each group.

The universal vowel production characteristics were obtained for alaryngeal and normal speech. The acoustic characteristics presented in Fig. 4 and 5, match the theoretical frequencies of Polish vowels formants from [3]. It can be seen, that for the normal speech the vowels subspaces in F1 and F2 formants dimensions can be linearly separated (Fig. 4). The relative positions of formant frequencies were maintained for alaryngeal voice (Fig. 5).

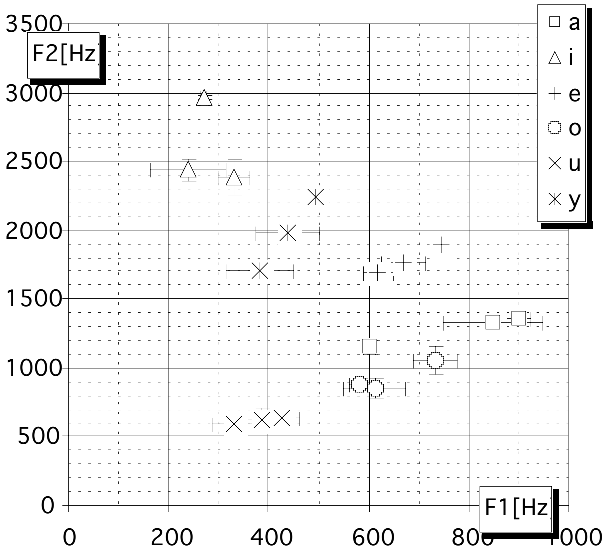

Fig 4. Vowels in normal speech (Polish) in the F1 and F2 formants dimensions

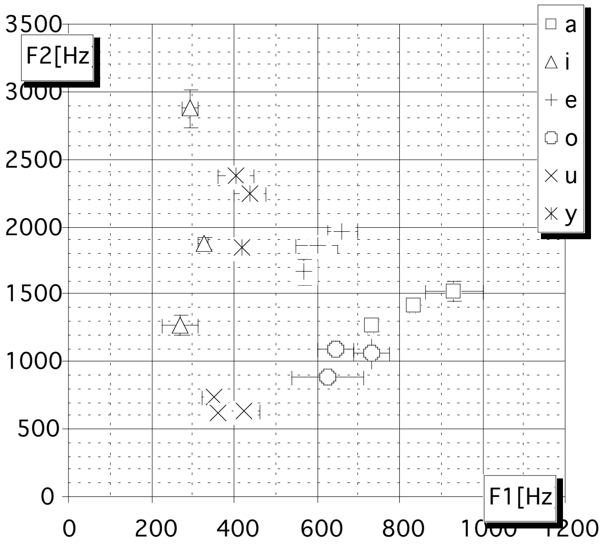

Fig 5. Vowels in esophageal speech (Polish) in the F1 and F2 formants dimensions

The Fig. 6 shows how the pathology affects the speech spectrum. For the patients, who haven't learned the esophageal voice, it is impossible to recognize vowels by two first formants. Moreover, it has been observed that the formants of vowels in the silent mouthing are more dispersed, unclear and less stable than in alaryngeal voice. Significant deviations of temporal formant frequencies are seen on Fig. 6.

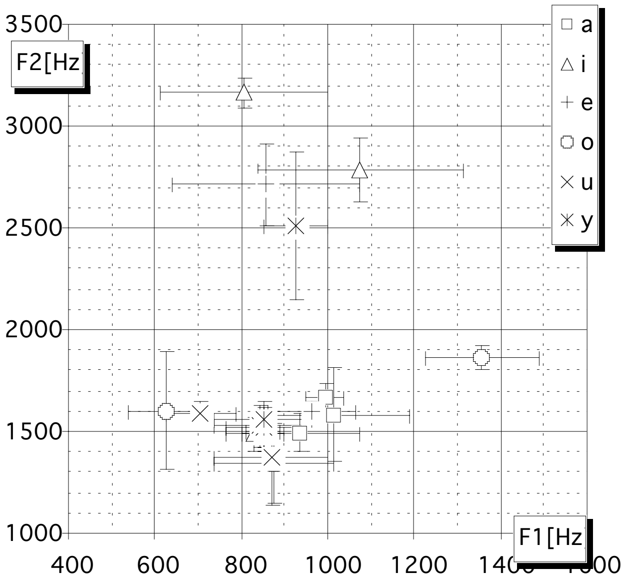

Fig 6. Vowels in silent mouthing speech (Polish) in the F1 and F2 formants dimensions

The figures with objective data prove how important the rehabilitation and learning an esophageal voice is. The pronunciation of vowels causes the biggest problems to laryngectomees. Vowels should be articulated with the use of low fundamental frequency. Patients should then learn how to use the source of phonation alternative to laryngeal voice, to acquire useful intelligible voice production.

IV. Discussion

Initial tests' results show the field for the future work on improving pathological speech with the computer methods. A linear or nonlinear separation methods, e.g. neural networks or SVM can be used for vowel recognition. However, we are not able to recognize silent mouthing speech based on two first formants; therefore we use other parameters, e.g. lips and jaw expression from image analysis.

V. Conclusion

It has been presented that formant frequencies equivalent to vocal tract coefficients are very sensitive to pathology of speech organs. This indicates that formants are objective descriptors for evaluation of rehabilitation after the laryngeal surgery and speech intelligibility. Our approach is developed for efficient and accurate tracking formants from the smooth AR spectrum. Eliminating noise and fundamental frequency with its harmonics due to the use of adaptive algorithm allows extracting formants in a simple way. The computer program on formant extraction can be used for objective analysis of vicarious voice of laryngectomees.

References

[1] B. Chen and P.C. Loizou, "Formant Frequency Estimation in Noise*"* in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing,* vol. I, Montreal, 2004, pp.581-584.

[2] L. J. Lee, H. Attias, L. Deng and A. Acero, "A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances" in *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, 2004.

[3] R. Tadeusiewicz, *Sygna_ mowy,* 1st ed., Warsaw: Wydawnictwa Komunikacji i Lacznosci, 1988, pp.158-186.

[4] R. Goldberg and L. Riek, Chapter 3, 4 in *A Practical Handbook of Speech Coders*, Boca Raton: CRC Press, 2000.

[5] T. Zielinski, *Od teorii do cyfrowego przetwarzania sygna_ów,* Krakow: Wydzial EAIiE AGH, 2002.

[6] L. Rutkowski, *Filtry adaptacyjne i adaptacyjne przetwarzanie sygna_ow: teoria i zastosowania,* Warsaw: Wydawnictwa Naukowo Techniczne, 1994, pp.49-85

[7] S. Saito, *Speech Science and Technology,* Tokyo: Ohmsha, 1992, pp. 63-94.

[8] Z. Zyszkowski, *Podstawy elektroakustyki,* 3 rd ed., Warsaw: Wydawnictwa Naukowo Techniczne, 1984, pp.218-277

[9] S. Haykin, *Adaptive filter theory*, Englewood Cliffs: Prentice Hall, 1991, pp. 477-485.

[10] R. Christensen, W. Strong and P. Palmer "A comparison of three methods of extracting resonance information from predictor-coefficient coded speech" in *IEEE Trans. Acoust., Speech, and Sig. Proc. ASSP-24(1),* 1976, pp. 8-14.