ARCHIVES OF ACOUSTICS
29, 1, 35–43 (2004)

DISCRETE COSINE TRANSFORM BASED DE-NOISING OF GLOTTAL PULSES

C. PASTIADIS and G. PAPANIKOLAOU

Laboratory of Electroacoustics
Electrical and Computer Engineering Dept.
Polytechnic School
Aristotle University of Thessaloniki, Greece
e-mail: pastiadi@egnatia.ee.auth.gr

(*received 2 September 2003; accepted 25 November 2003*)

Reliable estimates of the glottal function are of major importance in speech/voice processing for the characterization of voicing conditions, description of various phonation types, and identification of their parameters. This paper presents a new method for de-noising glottal wavelets (Differentiated Glottal Volume Velocity Pulses) and separation of the noise component, based on an approximation of their Discrete Cosine Transform as a sum of Exponentially Damped Sinusoids. The identification of the Exponentials' parameters leads to convenient estimation of "clean" glottal wavelets and thus separation of noise disturbances. The method is compared to standard Low-pass filtering and Wavelet de-noising using Monte Carlo simulations on synthetic Liljencrants–Fant glottal pulse models. As shown, the method supercedes for lower SNRs. Moreover, the method does not require exact determination of control parameters thus offering ease of implementation.

1. Introduction

The estimation of parameters of glottal pulses following inverse filtering of speech signals, requires removal of disturbances due to turbulent flow or other perturbation phenomena in the larynx region [1, 3]. Although an exact characterization of these disturbances and a clear distinction between pure glottal excitation and "noisy" components may not be easily qualified, the spectral content of these disturbances usually expands over wider frequency ranges than the underlying glottal function, despite any possible spectral concentration at some narrower bands [4–7, 16].

Speech inverse filtering has been extensively analyzed in the past for the estimation of the glottal excitation for voiced speech, while the use of methods based on Exponentially Damped Sinusoids approaches has recently offered better performance compared to classical Linear Prediction [16–19].

Generally speaking, signal de-noising has been widely addressed with many approaches such as the spectral separation of desired components and disturbances, the

use of various psychoacoustical criteria [10, 21], implementation of higher-order statistics, exploitation of wavelet transform properties [2, 8], etc.

A different approach to glottal pulse de-noising is presented in this work. Taking into account the transient profile of the differentiated glottal volume velocity signal, we propose a de-noising algorithm based on the identification of transient components in a noisy glottal period. Such an idea has been successfully employed for the recovery of lung sounds using the Wavelet transform [8]. In addition, some new methods for the identification of the attack portion of musical sounds from the rest of their "steady" part have recently been proposed [23]. These methods are based upon the use of the Cosine Transform, which virtually maps timings of occurrences onto frequencies of cosine functions. The method presented in this paper exploits such transformations and attempts retrieval of glottal pulses from respective noisy observations, using estimates of parameters of noisy exponentials in the Discrete Cosine Transform domain, and finally back transformation into the time domain.

The paper is organized as follows. Section 2 describes major properties of the Discrete Cosine Transform (DCT) and its characteristics for typical glottal pulses are introduced. The identification of Exponential components in the DCT is also established. Section 3 presents a Linear Prediction framework for the determination of the order of a Damped Exponentials mixture approximation of a glottal pulse's DCT. Based on such an approximation and the aforementioned properties of DCT, Sec. 4 introduces the Discrete Cosine Transform based Glottal Pulse Estimation algorithm. The DCT–GPE algorithm's de-noising efficiency is explored in Sec. 5 using Monte-Carlo simulations and a comparative study with standard Low-Pass filtering and Wavelet De-noising is also performed. Finally, Sec. 6 summarizes the aims of the work, concludes with major advantages of the proposed algorithm and discusses possible applications of such an approach.

2. DCT of a glottal pulse

The Discrete Cosine Transform [14] analyzes a sequence $x(n)$ using real orthogonal base functions such as the cosine. In this paper we deal with the so called "form II" of the transformation, which is defined as

$$\begin{aligned} X(k) &= \sqrt{\frac{2}{N}}\beta(k)\sum_{n=0}^{N-1} x(n)\cos\left(\frac{(2n+1)k\pi}{2N}\right), \qquad 0 \le k \le N-1, \\ \beta(k) &= \begin{cases} \dfrac{1}{\sqrt{2}}, & k = 0, \\ 1, & k = 1, ..., N-1. \end{cases} \end{aligned} \tag{1}$$

From Eq. (1) it follows that a Dirac delta in $x(n)$ at time n_i is represented in DCT domain as a cosine with frequency $(2n_i+1)/4N$, where N is the $x(n)$ sequence length.

Respectively, the DCT of a more complex pulse-like signal, such as the glottal pulse, consists of a sum of cosine functions with appropriate frequencies and amplitudes (Fig. 1).

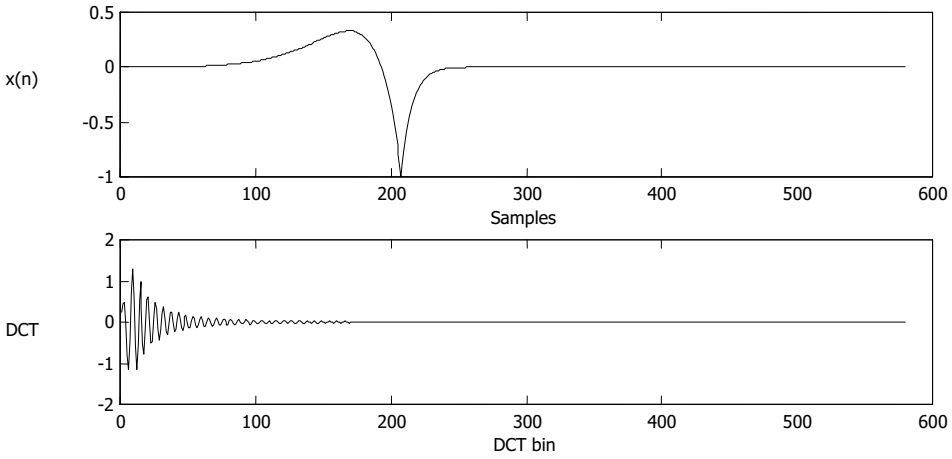

Fig. 1. A typical glottal pulse and its DCT.

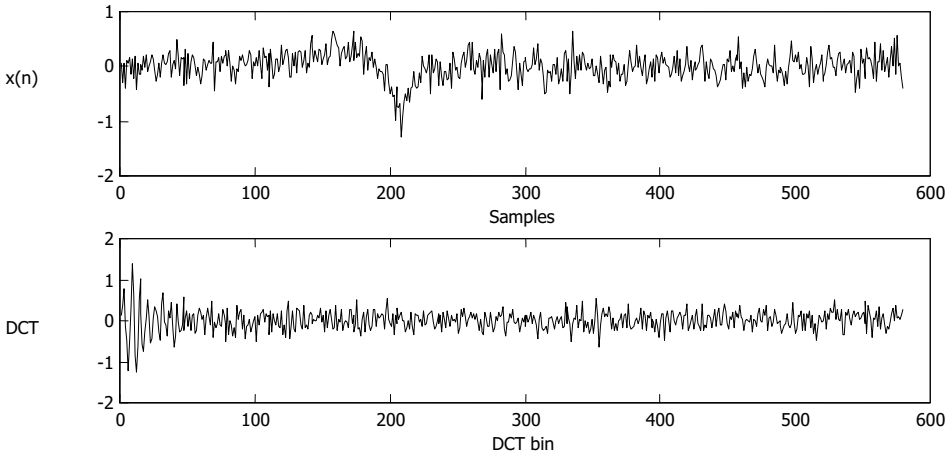

Fig. 2. A typical noisy glottal pulse and respective DCT.

As observed, the DCT of a glottal pulse is "damped" for higher values of the DCT bin. Such a situation may be attributed to the fact that since the glottal pulse itself may be regarded as a sequence of nearby delta-like pulses (mainly at the region of maximum stimulation), its DCT consists of cosine functions with frequencies very close to each other. As a result, DCTs of noisy glottal pulses exhibit lower SNRs at higher values of k (Fig. 2). In addition, the glottal noise component manifests itself in DCT domain as an additive noise-like disturbance spanning the whole DCT bins range.

3. Linear prediction models of a glottal pulse's DCT

As previously stated, the DCT is actually a sum of appropriate cosine functions. We also presented DCT patterns of typical noise-free and noisy glottal pulses (Differentiated Glottal Volume Velocity). The above, and in conjunction with the fact that a harmonic function may be described using an Auto-Regressive linear prediction model, lead to the consideration of such a model's order for a description of DCT with sufficiently low error, since the energy of a glottal pulse is mainly localised in a glottal period's region. Stated differently, we seek for the order p of an auto-regressive model such that the DCT $X_{\mathrm{DCT}}(k)$ of a glottal pulse $x_p(n)$ can be described as:

$$X_{\mathrm{DCT}}(k) = -\sum_{i=1}^{p} a_i X_{\mathrm{DCT}}(k-i) + e_{\mathrm{DCT}}(k), \qquad k = p+1, ..., N-1. \quad (2)$$

The estimation of a_i is based on minimization of the Mean Square Error $\overline{e^2_{\mathrm{DCT}}(k)}$.

Figure 3 shows the above MSE $\overline{e^2_{\mathrm{DCT}}(k)}$, together with the Akaike Information Criterion (AIC) [11, 12] for linear prediction as a function of the auto-regressive model order p.

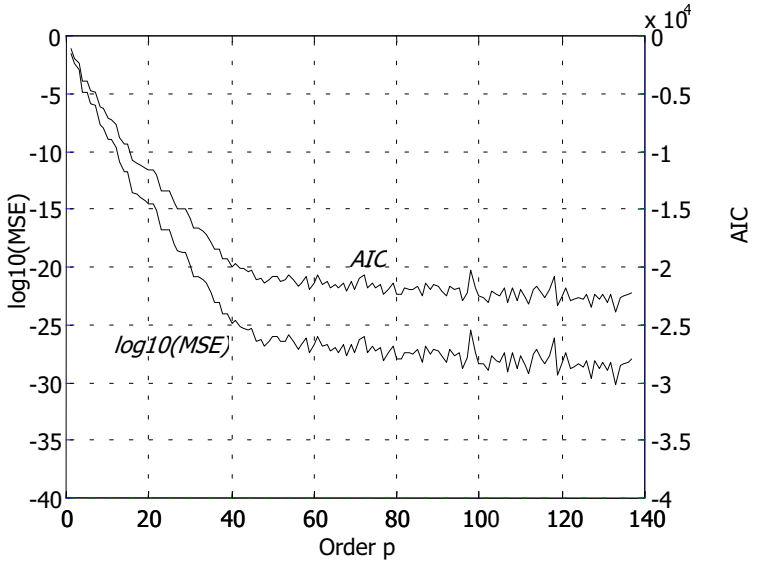

Fig. 3. MSE and AIC for linear prediction of a Glottal Pulses's DCT.

It can be observed that both MSE and AIC tend to stabilize for $p \geq$ 50–60, while any further increase of the model order does not contribute to any remarkably dramatic decrease. This observation leads to a possible description of a glottal pulse's DCT using an AR model of relatively low order ($\sim$ 50–60). From another point of view, the DCT of a glottal pulse may be fairly described using a reduced order sum of cosine functions

($\sim$ 25–30). It must be pointed that the orders determined by the previous approach actually depend on the sampling frequency for the glottal pulse in discrete time. The values given above are obtained for 48 kHz sampling frequency, while lower sampling rates would lead to lower order values. The determination of the parameters of such a model may follow various Exponentially Damped Sinusoids estimation frameworks [9, 15, 16, 20, 22]. The above mentioned approach is the foundation for the proposal of the de-noising algorithm that follows.

4. De-noising of glottal pulses: DCT–GPE algorithm (Discrete Cosine Transform-Glottal Pulse Estimation)

Based on the previous analysis, the proposed de-noising algorithm consists of the following steps:

1. From a noisy glottal pulse $g(n) = g_x(n) + g_n(n)$, where $g_x(n)$ is the noise-free pulse and $g_n(n)$ is the noise component, we form its DCT $G(k) = \text{DCT}(g(n))$.
2. From the "noisy" $G(k)$ we estimate the linear prediction coefficients a_i of an initial model of lower order ($p \approx 15$). These coefficients form a linear prediction polynomial with roots designated as z_i, which may be found using various schemas for the estimation of parameters of noisy Exponentially Damped Sinusoids such as the Matrix Pencil, Kumaresan-Tufts, Modified Higher Order Statistics method (M.H.O.S.), etc. [18]. Initial estimates of z_is may be supported using band pass filtering around the DCT's center frequency relating to the glottal pulse's main excitation location.
3. An estimate of the DCT of the noise-free signal $g_x(n)$ may be found as:

$$\widehat{G}_x(k) = \begin{cases} -\sum_{i=1}^{p} a_i \widehat{G}_x(k-i), & k = p+1, ..., N-1, \\ G(k), & k = 1, ...p. \end{cases} \tag{3}$$

   In this way, we estimate $\widehat{G}_x(k)$ as a lower order cosine approximation through an autoregressive model, with initial conditions $\widehat{G}_x(k) = G(k), k = 1...p$ considering that the local SNR in the DCT domain is higher for small k, since $G_x(k)$ exhibits damping as k increases. The selection of a lower initial order p aims to the identification of the main excitation of the DCT, where the implied SNR is higher.
4. An estimate of the noise component's $g_n(n)$ DCT is:

$$\widehat{G}_n(k) = G(k) - \widehat{G}_x(k). \tag{4}$$

5. Using low order ($\sim$5) AR spectral estimation for $\widehat{G}_n(k)$, we form the PSD (Power Spectral Density) of $\widehat{G}_n(k)$ as $\widehat{G}_n(f)$, and employing amplitude spec-

tral subtraction we get an estimate of the "de-noised" DCT as:

$$\widehat{G}_{\text{denoised}}(k) = IFT\left[\left||G(f)| - \left[\widehat{G}_n(f)\right]^{1/2}\right| \angle \arg(G(f))\right]. \qquad (5)$$

6. Setting $G(k) = \widehat{G}_{\text{denoised}}(k)$, repeat steps 2–6 with increasing p in each iteration, until $\widehat{G}_n^2(k)$ converges.
7. Finally, we obtain an estimate of the noise-free glottal pulse using an Inverse DCT, namely $\widehat{g}_x(n) = \text{IDCT}(\widehat{G}_x(k))$.

5. Application of DCT–GPE on noisy glottal wavelets and comparative assessment

DCT–GPE efficiency is compared to two other major de-noising schemes, namely Wavelet de-noising and linear low-pass filtering for synthetic noisy glottal pulses.

Synthetic noisy glottal pulses ($F_s = 48$ kHz) consist of a typical noise-free Lilljencrants–Fant's model glottal pulse (LF model) and additive noise with spectral characteristics as described in [5, 16], for various glottal SNRs. It must be noted that even very low glottal SNRs may manifest themselves at higher uttered voice SNRs due to the selectivity of the vocal tract all-pole filter.

The DCT-GPE algorithm is compared to Wavelet de-noising using the "db9" (Daubechies-9) wavelet at 6-th level and soft thresholding. The Daubechies-9 wavelet and 6-th level were selected as the best choices after a comparative observation of various wavelet families and analysis orders. Moreover, the low-pass spectral profile of typical glottal pulses motivates the use of linear low-pass filtering for the removal of wide-band glottal noise. The use of optimal low-pass filters with predetermined parameters (order and cutoff frequency) following an investigation for minimization of "clean" glottal wavelet's estimation error is also employed in our comparative study.

Our study also includes the possible effect of various OQ-Open Quotient values, since the glottal SNR increases for constant rms of the noise component and higher OQ, thus resulting at different local SNRs around the glottal pulse's maxima region. For this reason two different OQ values (40% and 80%) are included in our study.

Table 1 summarizes the findings for the mean rms of de-noising error $e(n) = g_x(n) - \widehat{g}_x(n)$, over 500 Monte–Carlo runs for each glottal SNR and $\text{OQ} = 80\%$.

Table 2 presents the simulations' findings for $\text{OQ} = 40\%$.

As it can be seen, the DCT–GPE algorithm is superior to Wavelet de-noising (up to ~ 3.5 dB) for decreasing SNR and especially at higher OQ values (as it could be expected). Moreover, the use of low-pass filtering may offer better results than Wavelet de-noising for the specified parameter values. Such a finding may be attributed to the fact that the low-pass filter parameters' selection approximates an optimal filter's efficiency [13], relying to the SNR and PSDs of the noisy signal and noise component, which however are *a priori* unknown. It must be once more pointed that the low-pass filter's parameters are selected among various combinations using a e_{rms}^{lpf} minimization criterion.

Table 1. RMS of de-noising error under DCT–GPE, Wavelet De-noising, and Linear Low-Pass Filtering for $\mathrm{OQ} = 80\%$.

SNR(dB) $[\mathrm{OQ} = 80\%]$	$e_{rms}^{\mathrm{DCT-GPE}}$	$e_{rms}^{\mathrm{wavelet}}$	e_{rms}^{lpf}, filter order, f_c (Hz)	$20\log_{10}\frac{e_{rms}^{\mathrm{DCT-GPE}}}{e_{rms}^{\mathrm{wavelet}}}$	$20\log_{10}\frac{e_{rms}^{\mathrm{DCT-GPE}}}{e_{rms}^{lpf}}$
15	0.0236	0.0227	0.0231, 7, 1100	0.3	0.18
10	0.0335	0.0407	0.0385, 10, 1100	-1.7	-1.2
5	0.0493	0.0740	0.0662, 11, 500	-3.53	-2.56

Table 2. RMS of de-noising error under DCT-GPE, Wavelet De-noising, and Linear Low-Pass Filtering for $\mathrm{OQ} = 40\%$.

SNR(dB) $[\mathrm{OQ} = 40\%]$	$e_{rms}^{\mathrm{DCT-GPE}}$	$e_{rms}^{\mathrm{wavelet}}$	e_{rms}^{lpf}, filter order, f_c (Hz)	$20\log_{10}\frac{e_{rms}^{\mathrm{DCT-GPE}}}{e_{rms}^{\mathrm{wavelet}}}$	$20\log_{10}\frac{e_{rms}^{\mathrm{DCT-GPE}}}{e_{rms}^{lpf}}$
15	0.0143	0.0129	0.0142, 6, 1700	0.89	0.06
10	0.0256	0.0283	0.0297, 12, 500	-0.9	-1.29
5	0.0428	0.0499	0.046, 13, 200	-1.35	-0.62

Figure 4 shows an implementation of the DCT–GPE and Wavelet de-noising algorithms on a typical synthetic glottal pulse extracted from the executed simulations. The SNR is 10 dB.

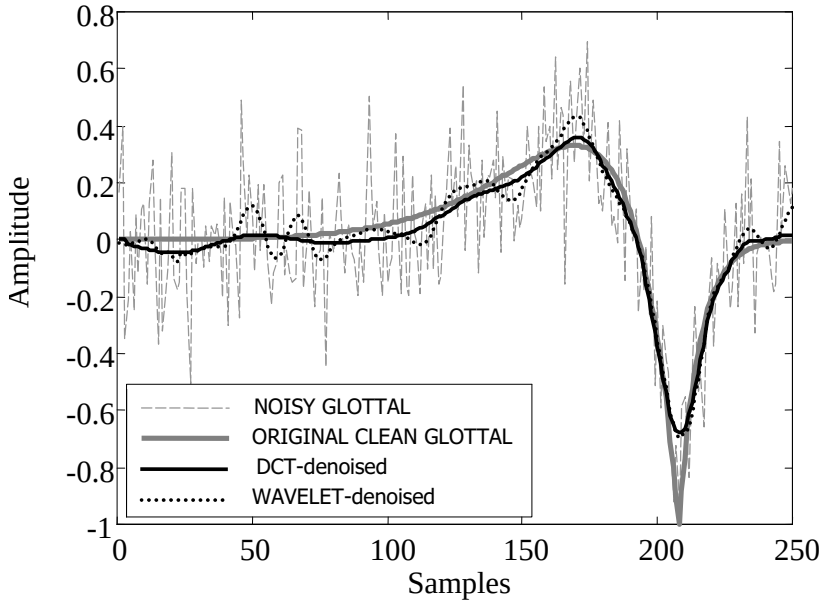

Fig. 4. Noisy glottal pulse de-noising following the proposed DCT–GPE algorithm, Wavelet de-noising, together with "clean" and "noisy" initial pulses. $\mathrm{SNR} = 10$ dB.

As it can be observed, both the Wavelet-based de-noising and the DCT–GPE algorithm perform similarly at the region of maximum excitation (~ 210 samples). The sharpness of excitation is retained fairly well. However, the DCT–GPE performs much better at the rest of the glottal pulse showing a remarkably good fit to the original "clean" signal. Instead, the Wavelet de-noising approach exhibits a moderate ripple, spanning almost over the whole pulse duration. The presence of such a ripple may cause substantial ambiguity for the identification of possible secondary excitation regions during a glottal cycle, especially in pathological voices. Similar effects characterize the performance of the examined algorithms for the majority of the Monte–Carlo simulation runs.

After the estimation of the noise-free component, a direct estimation of the noise component is also possible, together with its spectral characteristics and Signal-to-Noise Ratio.

6. Conclusions

A new algorithm (Discrete Cosine Transform-Glottal Pulse Estimation) is introduced for the estimation of the noise component of noisy glottal pulses, based on the use of Discrete Cosine Transform. Exponentially Damped Sinusoids estimation in the DCT domain is facilitated for the identification of the underlying clean glottal wavelets. The proposed algorithm offers higher efficiency for the estimation of the noise-free glottal pulse over two standard de-noising methods that employ Wavelet analysis and linear low-pass filtering. The method does not require precise parameter tuning, thus becoming simple-to-use. The estimation of the glottal noise component may offer new possibilities to the determination of both local and overall glottal SNR, which may be beneficial for acoustic analysis of laryngeal pathologies. At the moment the method is employed and tested in a voice processing environment used for analysis of healthy and pathological voice recordings. Additionally, the method may be used in a variety of applications that require identification of "peaky" or "spiky" components in time or other domains. Such a field is that of x-ray crystallography, where the method is currently being clarified for use.

References

[1] P. BADIN, E. CASTELLI, Y.P.T. NGOC, *Acoustic transfer functions for vowels and consonants*, Speech Maps (Esprit/br No 6975), Appendix D, 1–19 (1992).

[2] S. BURRUS, R.A. GOPINATH, H. GUO, *Introduction to wavelets and wavelet transforms*, Prentice Hall, New Jersey 1998.

[3] E. CASTELLI, P. BADIN, *Time and frequency domain acoustic models of the vocal tract*, Speech Maps (Esprit/br No 6975), Appendix B, 1-25 (1992).

[4] D.G. CHILDERS, C.K. LEE, *Vocal quality factors: analysis, synthesis, and perception*, JASA, **90**, 5, 2394–2410, November (1991).

[5] D.G. CHILDERS, *Speech processing and synthesis toolboxes*, John Wiley & Sons, New York 2000.

[6] C.H. COKER, M.H. KRANE, B.Y. REIS, R.A. KUBLI, *Search for unexplored effects in speech production*, Proc. ICSLP, 1996.

[7] P.R. COOK, *Identification of control parameters in an articulatory vocal tract model, with applications to the synthesis of singing*, Ph.D. Thesis, Princeton, September 1991.

[8] L. HADJILEONTIADIS, C. LIATSOS, C. MAVROGIANNIS, T. ROKKAS, S. PANAS, *Enhancement of bowel sounds by wavelet-based filtering*, IEEE Trans. Biomedical Eng., **47**, 7, 1–11, July (2000).

[9] Y. HUA, T. SARKAR, *Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise*, IEEE Trans ASSP, **38**, 5, 814–824, May (1990).

[10] G. KALLIRIS, *New techniques for restoration of speech and music*, Ph.D. Dissertation, Aristotle University of Thessaloniki, 1995.

[11] S. KAY, *Modern spectral estimation*, Prentice Hall PTR, Signal Processing series, New Jersey 1988.

[12] L. LJUNG, *System identification*, Prentice Hall PTR, Signal Processing series, New Jersey 1987.

[13] P.A. NELSON, S.J. ELLIOTT, *Active control of sound*, Academic Press, London 1992.

[14] A. OPPENHEIM, R. SCHAFER, J. BUCK, *Discrete time signal processing*, Prentice Hall PTR, Signal Processing Series, New Jersey 1998.

[15] C.K. PAPADOPOULOS, C.L. NIKIAS, *Parameter estimation of exponentially damped sinusoids using higher order statistics*, IEEE Trans. ASSP, **38**, 8, 1424–1435, August (1990).

[16] C. PASTIADIS, *Contemporary voice analysis techniques for the speech impaired*, Ph.D. Dissertation, Aristotle University of Thessaloniki, 2002.

[17] C. PASTIADIS, G. PAPANIKOLAOU, *A preliminary study on greek esophageal speech and a method for quality and intelligibility enhancement*, Archives of Acoustics, **24**, 1, 25–38 (1999).

[18] C. PASTIADIS, G. PAPANIKOLAOU, *Higher-order statistics based inverse filtering for analysis of esophageal voice production*, 104-th Audio Engineering Society Convention, May 1998.

[19] C. PASTIADIS, A. PRINTZA, S. METAXAS, I. DANIILIDIS, G. PAPANIKOLAOU, *Acoustic voice analysis in the evaluation of vocal fold polyps*, Proc. AFEA 2001, Athens, June 2001.

[20] B. PORAT, B. FRIEDLANDER, *A modification of the Kumaresan–Tufts method for estimating rational impulse response*, IEEE Trans. ASSP, **34**, 5, 1336–1338, October (1986).

[21] D. TSOUKALAS-STATHAKIS, *Noise subtraction from speech and noise signals*, Ph.D. Dissertation, University of Patras, 1997.

[22] S. VAN HUFFEL, H. CHEN, C. DECANNIERE, P. VAN HECKE, *Total least squares based algorithm for time-domain NMR data fitting*, ESAT Laboratory-Katholieke Universiteit Leuven 1994.

[23] T. VERMA, S. LEVINE, T. MENG, *Transient modeling synthesis: a flexible analysis/synthesis tool for transient signals*, Proc. ICMC 97, Thessaloniki 1997.