ARCHIVES OF ACOUSTICS
24, 1, 25–38 (1999)

A PRELIMINARY STUDY ON GREEK ESOPHAGEAL SPEECH AND A METHOD FOR QUALITY AND INTELLIGIBILITY ENHANCEMENT

C. PASTIADIS and G. PAPANIKOLAOU

Laboratory of Electroacoustics,
Aristotle University of Thessaloniki
(GREECE)

The present work is a preliminary study on Greek esophageal speech and is mainly concerned with the investigation of major features such as pitch, formant frequencies, and speech power envelopes. The implementation in esophageal speech of various well-known techniques for normal voice analysis is overviewed. An improved method for resynthesizing voiced sounds (such as vowels or nasal consonants) by convolution of an ARMA estimate of the speaker's vocal tract impulse response and a periodic glottal waveform is proposed as a tool for voice quality enhancement. Fundamental frequency values were confirmed to be close to previous works' findings. F1 and F2 formant alterations due to laryngectomy were not detected compared to normal speech values. However, speech power envelopes tended to be flatter as the speaker's training stage was higher. The proposed method for speech enhancement proved able enough to preserve speaker characteristics and provide cues for higher quality reproduction of vowels as well as nasals.

1. Introduction

Esophageal speech is produced by laryngectomized people who utter by expelling air constricted in their esophagus. The expelled air forces the cricopharyngeal cartilage to oscillate in a manner that imitates the vocal folds' operation. Although proper training may help esophageal speakers utter intelligibly enough, a severe degradation of voice quality after laryngectomy usually occurs. The voice quality of esophageal speech may be described as harsh, rough and low. The amount of aperiodicity is high and the voice is often very noisy.

Esophageal speech features have been investigated in the past. A part of previous works focused on pitch and intensity characteristics and their perceptual aspects [11, 12, 18, 19, 24].

Efforts for esophageal speech quality and intelligibility enhancement have also been reported [1, 2, 9, 16, 17, 23].

The present work is a preliminary study on Greek esophageal speech and is mainly concerned with the investigation of major features such as pitch (or fundamental frequency F0), formant frequencies, and speech power envelopes. The implementation of various normal voice analysis techniques in esophageal speech is overviewed [14]. An

improved method for resynthesizing voiced sounds (such as vowels or nasal consonants) is proposed as a tool for voice quality enhancement [14].

2. Subjects and recordings

Nine male alaryngeal speakers were used as subjects for uttering various CV sequences and full sentences. CV sequences comprised of all Greek vowels and stop consonants.

Speakers were selected among groups of various grades of speech production training and had no treatment prior to voice recording.

Utterances were spoken at normal rate and level and were recorded on a DAT recorder through an electret condenser microphone placed at a distance of approximately 20 cm from the speaker's mouth. A sampling rate of 11 kHz was adopted for further computer processing.

3. Investigation of speech features

Fundamental or pitch frequency, speech power envelope during phonations and formant frequencies are considered as some of the most important features for speech comprehension, training and clinical evaluation of voice. These features were investigated and the methods employed together with analysis results are presented.

3.1. F0 *investigation*

F0 or fundamental frequency was investigated for all nine speakers during vowel phonations in discrete CV contexts and isolated vowel segments through whole sentences.

Estimation of F0 was performed using well known methods for F0 extraction on normal speech, such as the Autocorrelation Method (biased and unbiased), the Center-Clipped Autocorrelation and 3-level Center-Clipped Autocorrelation methods, the Average Magnitude Difference Function method (AMDF), the Cepstrum and a so-called "Hubert– Envelope" Method.

3.1.1. The autocorrelation method. The autocorrelation method [6] identifies F0 or period of voiced speech by finding the lag at which the autocorrelation function of the speech signal is maximized, that is:

$$T_0 = \frac{1}{f_0} = \max_m \{\Phi(m)\} \cdot T_s \,, \tag{3.1}$$

where

$$\Phi(m) = \frac{1}{N} \sum_{n=0}^{N-1-|m|} s(n)s(n+|m|), \qquad m = 0, \ldots, N-1 \tag{3.2}$$

with $s(n)$ = speech signal, N = number of samples, T_s = sampling period.

We may also use the unbiased autocorrelation estimator:

$$\Phi(m) = \frac{1}{N-|m|} \sum_{n=0}^{N-1-|m|} s(n)s(n+|m|), \qquad |m| = 0, ..., N-1. \tag{3.3}$$

Estimation of the first maximum of the autocorrelation function using the biased and unbiased estimators is shown in Figs. 1b, 1c, respectively.

3.1.2. The center-clipped and 3-level center-clipped autocorrelation methods. Further improvement in the estimation of maxima of the autocorrelation function may be achieved using either center-clipping [21] or 3-level center-clipping on the speech signal:

$$s'(n) = \begin{cases} s(n) - C^+ & s(n) > C^+, \\ 0 & C^- \le s(n) \le C^+, \\ s(n) - C^- & s(n) < C^-, \end{cases} \tag{3.4}$$

$s'(n)$ = center-clipped speech signal, or

$$s'(n) = \begin{cases} 1 & s(n) > C^+, \\ 0 & C^- \le s(n) \le C^+, \\ -1 & s(n) < C^-, \end{cases} \tag{3.5}$$

$s'(n)$ = 3-level center-clipped speech signal and C^+, C^- are threshold values.

Estimated maxima of the autocorrelation tunction after center-clipping of the speech signal are shown in Fig. 1d.

3.1.3. The AMDF *(Average Magnitude Difference Function) method.* The AMDF method [7] tries to locate a strong minimum of the AMDF

$$\text{AMDF}(m) = \sum_n |s(n) - s(n+m)|, \tag{3.6}$$

$s(n)$ = speech signal.

For a strictly periodic speech signal the AMDF would take on a value of zero at $m = T_0/T_s$, $T_0 = 1/f_0$.

For quasi-periodic signals a strong minimum usually occurs at the period lag.

Results of this method on an esophageal speech sample are shown in Fig. 1e.

3.1.4. The cepstrum. This well known technique relies on the fact that time of maximum at the high-quefrency region of the cepstrum represents the period of the speech signal [6, 13]. Results of this method employed on esophageal speech are shown in Fig. 1f.

3.1.5. The "Hilbert-Envelope" method. This method [3] performs maximum-likelihood epoch determination as the basis for the estimation of glottal closure (epoch) instants (GCI) in normal voices and implements a Hilbert transformation for the improvement of its performance and reliability. It is also posed that the method may cover

most speech signals (even under noisy conditions). Since it is capable of estimating closure instants, it can sense period-to-period variations or nonstationary period variations within longer frames.

The method is employed on esophageal speech under the assumption that even in this kind of severely damaged vocal function there must exist a moment that a main pulse excites the whole nasopharyngeal system. Thus, the method tries to locate the maximally possible instants of main excitation.

The method initiates with the formation of a so-called Maximum-Likelihood Epoch Determination signal ("MLED-signal") which is proven to be a cross-correlation between the speech signal and the impulse response of the nasopharyngeal system filter due to an epoch, that is

$$\hat{f}(k) = \sum_{n=0}^{N-1} s(n+k)\hat{s}(n), \tag{3.7}$$

where $s(n)$ = speech signal, and

$$\hat{s}(n) = \begin{cases} G & n = 0, \\ -\sum_{i=1}^{p} a_i \hat{s}(n-i) & 0 < n \leq \infty, \\ 0 & \text{otherwise} \end{cases} \tag{3.8}$$

which is virtually the nasopharyngeal filter's impulse response obtained by AR modeling of the speech signal and a_i are the model's coefficients. The use of a high order AR modeling in esophageal speech is motivated by the fact that the glottal function is generally unknown and the nasopharyngeal system's function may include zeroes that should normally be represented by ARMA modeling. A selection of $p \geq 40$ was made in order to compromise between accuracy and computation time, although a value of at least $N/5$ where N = record length is suggested [10].

GCI's are identified as the time indices of maxima of the MLED signal. To reduce ambiguity in selection of maxima the MLED is multiplied by a Hubert–Envelope of itself,

$$\hat{f}(k) \cdot \hat{g}(k),$$

where

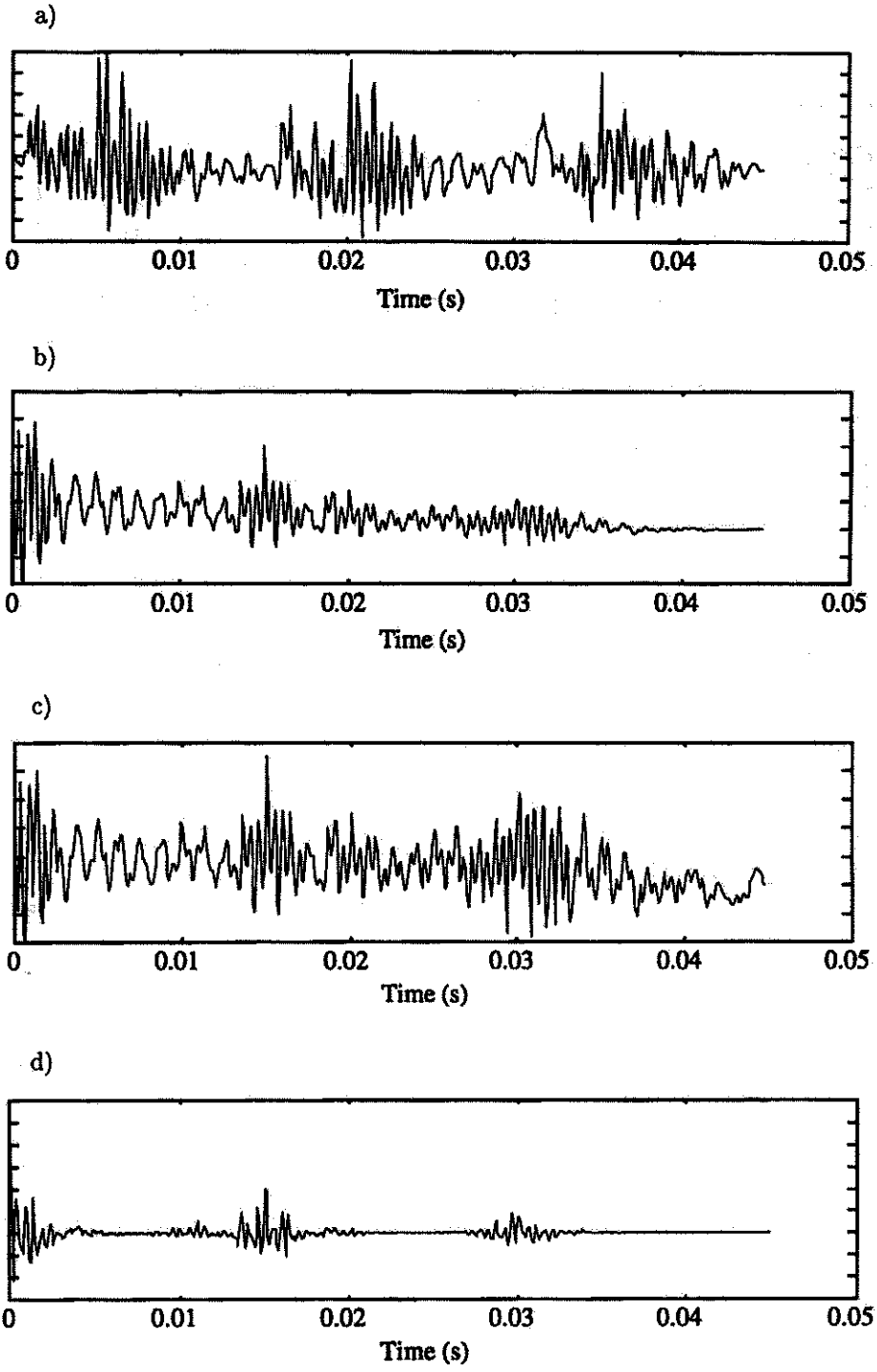
$$\hat{g}(k) = \sqrt{\left[\hat{f}^2(k) + \hat{f}_H^2(k)\right]}, \qquad \text{and} \qquad \hat{f}_H(k) = HT(\hat{f}(k)). \tag{3.9}$$

The Hilbert–Envelope proves able to emphasize the contrast between the main epoch pulse and other possible sub-pulses that indicate sub-optimal excitation instants.

The results of this algorithm are shown in Fig. 1g, where values of the MLED signal lower than a predetermined threshold were set to zero.

As it is observed, optimal instants of excitation are possible to locate in esophageal speech too.

Since this method of F0 estimation relies on the interpretation of fundamental frequency as the inverse of between-excitation period, it can provide information on


a)
0
0.01
0.02
0.03
0.04
0.05
Time (s)
b)
0
0.01
0.02
0.03
0.04
0.05
Time (s)
c)
0
0.01
0.02
0.03
0.04
0.05
Time (s)
d)
0
0.01
0.02
0.03
0.04
0.05
Time (s)

[Fig. 1]

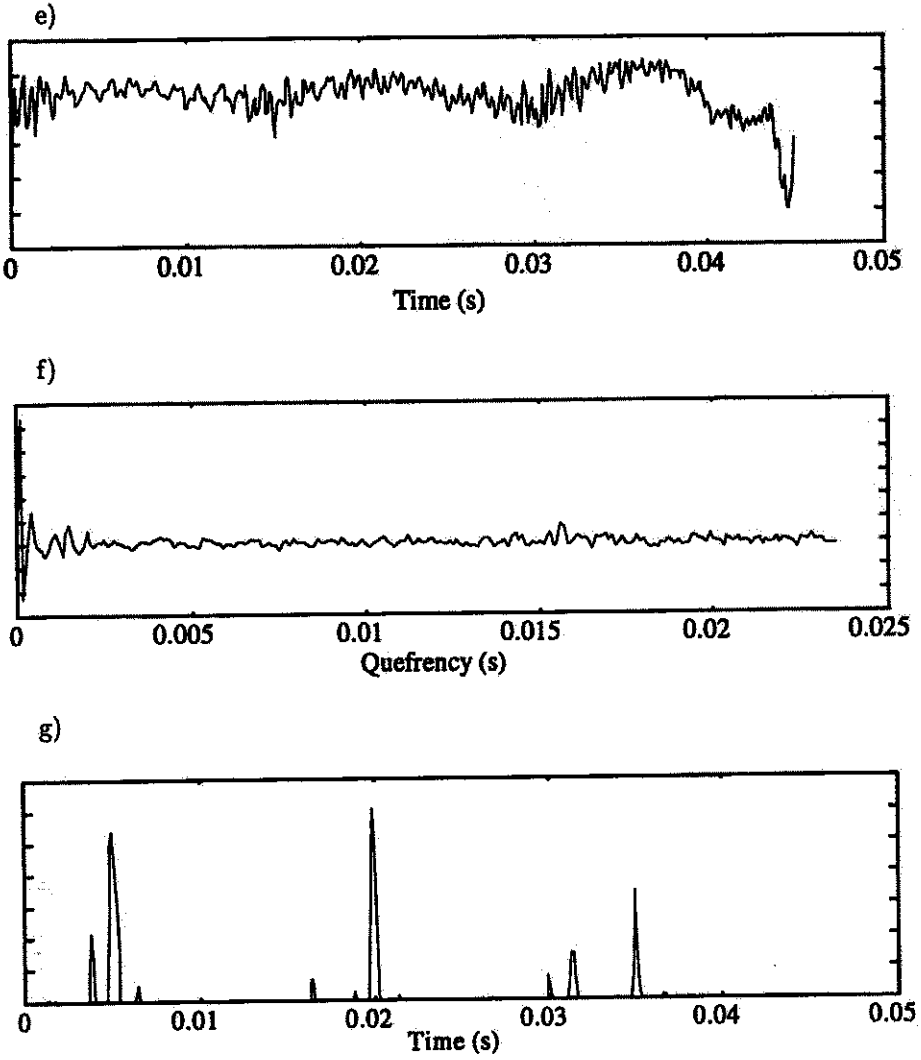


Fig. 1. Esophageal vowel /e/ (a), and fundamental period estimates using: the biased autocorrelation (b), the unbiased autocorrelation (c), the center-clipped autocorrelation (d), the AMDF (e), the Cepstrum (f) and the "Hilbert Envelope" (g) methods.

period-by-period F0, and thus it can be used for cycle-to-cycle estimation of glottal activity fluctuation (such as jitter or shimmer) [22]. The existence of various degrees of ambiguity in selection of GCI's is currently being investigated as a tool for the clinical evaluation of voice quality and/or training procedure's progress.

3.1.6. Results and discussion on F0 *investigation.* Though all the employed extraction methods are capable of estimating rough F0 values, ambiguity is lower when estimates are taken using the Center-Clipped Autocorrelation methods and/or the "Hilbert–

Envelope" method, which seem to provide more consistent F0 estimates to manually extracted ones. However, since the Autocorrelation-based methods are short-term average methods, they are not applicable to cycle-to-cycle variations investigation [22], whereas the "Hilbert–Envelope" method seems more appropriate.

Mean intra-speaker F0 (estimated using the Center-Clipped Autocorrelation method) values and standard deviation together with mean F0 values and standard deviation between all speakers are presented in Table 1.

Table 1. Intra-speaker F0 mean and standard deviation values together with inter-speaker mean and standard deviation values.

Speaker	Mean F0 (Hz)	S.D. (Hz)
1	57.5	9.0
2	64.9	9.8
3	86.4	19.9
4	88.4	19.9
5	73.8	15.1
6	82.2	19.0
7	89.9	18.4
8	43.3	9.0
9	68.2	12.1
inter-speaker Mean value	**72.7**	**14.7**
inter-speaker S.D.	**14.9**	**4.4**

As observed, mean F0 for all speakers is found at 72.7 Hz, which verifies previous works' findings that esophageal speech is about 1 octave lower than normal speech [11, 18, 19, 24].

3.2. Speech power envelope investigation

The slope of speech power envelope during phonations exhibits major perceptual and clinical interest [19], since it may provide information on voice dynamics and the speaker's training progress.

Speech power envelopes of /pa/ utterances from all speakers were obtained using a pitch-period wide integration window, and mean slopes of phonations were computed. Results are presented in Table 2.

Mean value between all speakers was found to be −86 dB/sec, with standard deviation 13.7 dB/sec.

Additional information about the stage of training of each one of the speakers showed a tendency of decrease in slopes (numerically higher slope values) with training past.

Table 2. Phonation slopes for the utterance /pa/.

Speaker	Mean phonation slope (dB/sec)
1	−81.3
2	−78.9
3	−78.8
4	−74.3
5	−92.9
6	−116.1
7	−70
8	−100
9	−81.3
Mean Value	−86
Standard Deviation	13.7

3.3. Vowels' F1, F2 *investigation*

First two formant frequencies (F1, F2) were investigated using LPC, for all 5 Greek vowels α, ε, ι, o, ov. As it is well known, Greek vowels differ from other languages' vowels in that they are not rounded and thus are displaced in the F1/F2 space [8].

Table 3 gives mean F1, F2 values for all speakers together with mean values of normal Greek speech.

Table 3. F1, F2 mean values for all Greek vowels for esophageal and normal speech [7].

	Esophageal speech		Normal speech	
Greek Vowel	F1 (Hz)	F2 (Hz)	F1 (Hz)	F2 (Hz)
α	732	1390	~ 700	~ 1300
ε	521	1750	~ 475	~ 1700
ι	385	1823	~ 300	~ 2000
o	510	992	~ 450	~ 850
ov	420	1095	~ 350	~ 900

A general coincidence between normal and esophageal speech formant frequencies values is observed, which proposes that formant extraction methods for speech recognition may be used in esophageal speech too.

4. A method for esophageal speech quality and intelligibility enhancement

As already stated, esophageal speech is severely degraded. Perceptual judgments of esophageal voice characterize it as harsh, rough and low. Generally, esophageal speakers are able to produce intonational contrasts but listeners do not readily perceive the variation.

Previous works on esophageal voice rehabilitation had followed both surgical and speech signal processing procedures. Among the most well known surgical methods is one that uses a valve prosthesis that permits air to flow from trachea to esophagus [191. In the signal processing domain, recent works report methods of spectral substitution of esophageal speech [1, 2], resynthesizing speech using LPC spectral estimation [16, 17, 23], and elimination of undesirable phenomena (such as injection noise) during voice production [9].

The method proposed in this work uses ARMA estimation on esophageal speech. Since in esophageal speech the excitation signal is generally unknown, an all-pole model alone may not accurately represent the speech production procedure. Moreover, a pure all-pole model of relatively low order may not be sufficient for the analysis of voiced consonantal sounds that include significant zeroes in their spectral envelope such as nasals (/m/, /n/). ARMA modeling allows for the extraction of major spectral envelope profiles by extracting both formants (poles) and anti-formants (zeroes) frequencies and bandwidths. Estimation of parameters of an ARMA model that describes esophageal speech production is performed using a Least Squares Modified Yule–Walker Equations (LSMYWE) approach [10]. Next, voiced speech resynthesizing is performed using convolution of the oronasal tract filter's impulse response (obtained with the use of ARMA modeling and selection of appropriate pole/zeroes pairs) with a waveform that represents normal voices' vocal fold vibratory function [20].

The proposed method's functional diagram is shown in Fig. 2.

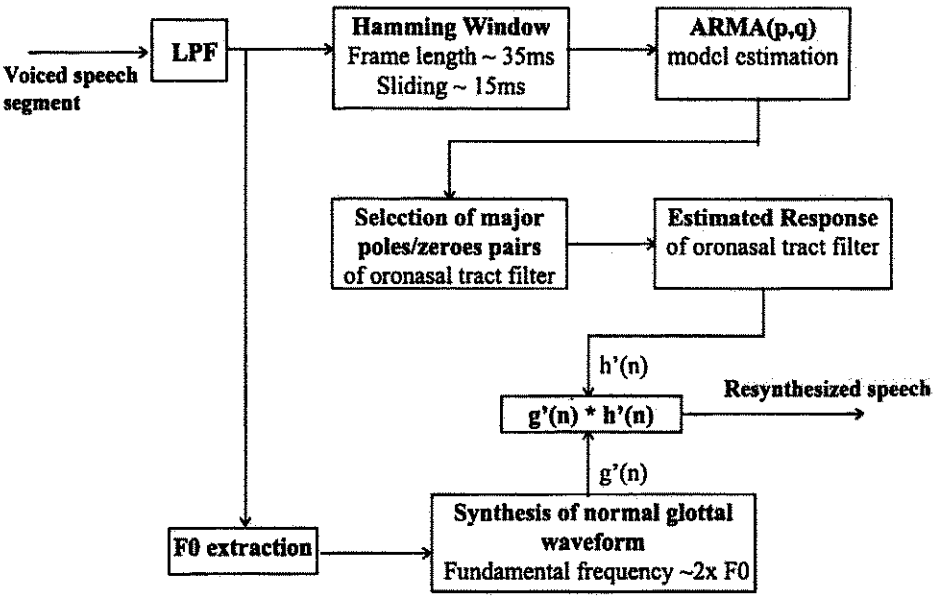


Fig. 2. Block diagram of the proposed method for voice quality enhancement.

As shown, a selection of voiced speech segments precedes the implementation of the method.

The ARMA modeled spectral envelope is of the form

$$S(z) = \frac{\sum_{i=0}^{q} b_i z^{-i}}{1 + \sum_{i=1}^{p} a_i z^{-i}}, \qquad b_0 = 1, \quad (p, q) = \text{ARMA model order.} \tag{4.1}$$

A $p \leq 16$ and $q \leq 10$ pair of values may be selected.

The selection of major poles and zeroes is made under the following assumptions:

a. Oronasal filter's poles are complex conjugate pairs with relatively high frequency-to-bandwidth ratios.

b. Significant zeroes also appear as complex conjugate pairs with relatively high frequency-to-bandwidth ratios, whereas real valued zeroes may reflect radiation and/or possible glottal waveform's spectral characteristics.

More specifically, the estimated filter's impulse response z-transform is:

$$V(z) = A \cdot \frac{\prod_{i=1}^{q'} (1 - z_i z^{-1})(1 - z_i^* z^{-1})}{\prod_{i=1}^{p'} (1 - p_i z^{-1})(1 - p_i^* z^{-1})}, \tag{4.2}$$

where z_i, p_i are selected zeroes and poles from the estimated ARMA model of esophageal speech according to the previous assumptions and $|p_i| < 1$.

Figure 3 shows the results of the LSMYWE analysis on steady-state portions of both the consonantal and vowelic regions of two original esophageal CV speech utterances /na/ and /me/. A 11 kHz sampling rate and a model order of $p = 12$ and $q = 5$ were selected. The analysis of the C part of the utterances seems to confirm previous works' findings on acoustic analysis of nasals. As it can be seen from the power spectra and the pole-zeros chart, a low frequency pole (nasal murmur) appears in the region of about 300 Hz for both /m/ and /n/. Moreover, the presence of side branch resonators (as the oral cavity in the case of nasals) introduces zeros in the spectrum of the uttered phoneme [15, 5]. Although these zeros are not very prominent, they produce a smoother energy distribution over different frequency ranges between /m/ and /n/ [15]. The zero introduced in the case of /n/ lies in the mid-frequency region (over 1 kHz), whereas a broader zero appears in the case of /m/ causing a more even energy distribution over the low-frequency region. In the case of the V utterances (/a/ and /e/), the formantic structure for each phoneme appears clearly, with formant values close to typical ones. The low-frequency nasal murmur is removed. Zeros in the estimated vocal tract response seem to exist close to mid-high formants, whereas the rest of them lie either on the real axis or they are of larger bandwidth; thus, they do not interfere significantly with the all-pole configuration of the vocal tract during vowel production.

A synthetic waveform that represents normal voice's glottal vibration is used for convolution with the previously estimated response. Triangular-like waveforms are preferred to pulse-like ones, since they seem to produce more intelligible synthetic speech

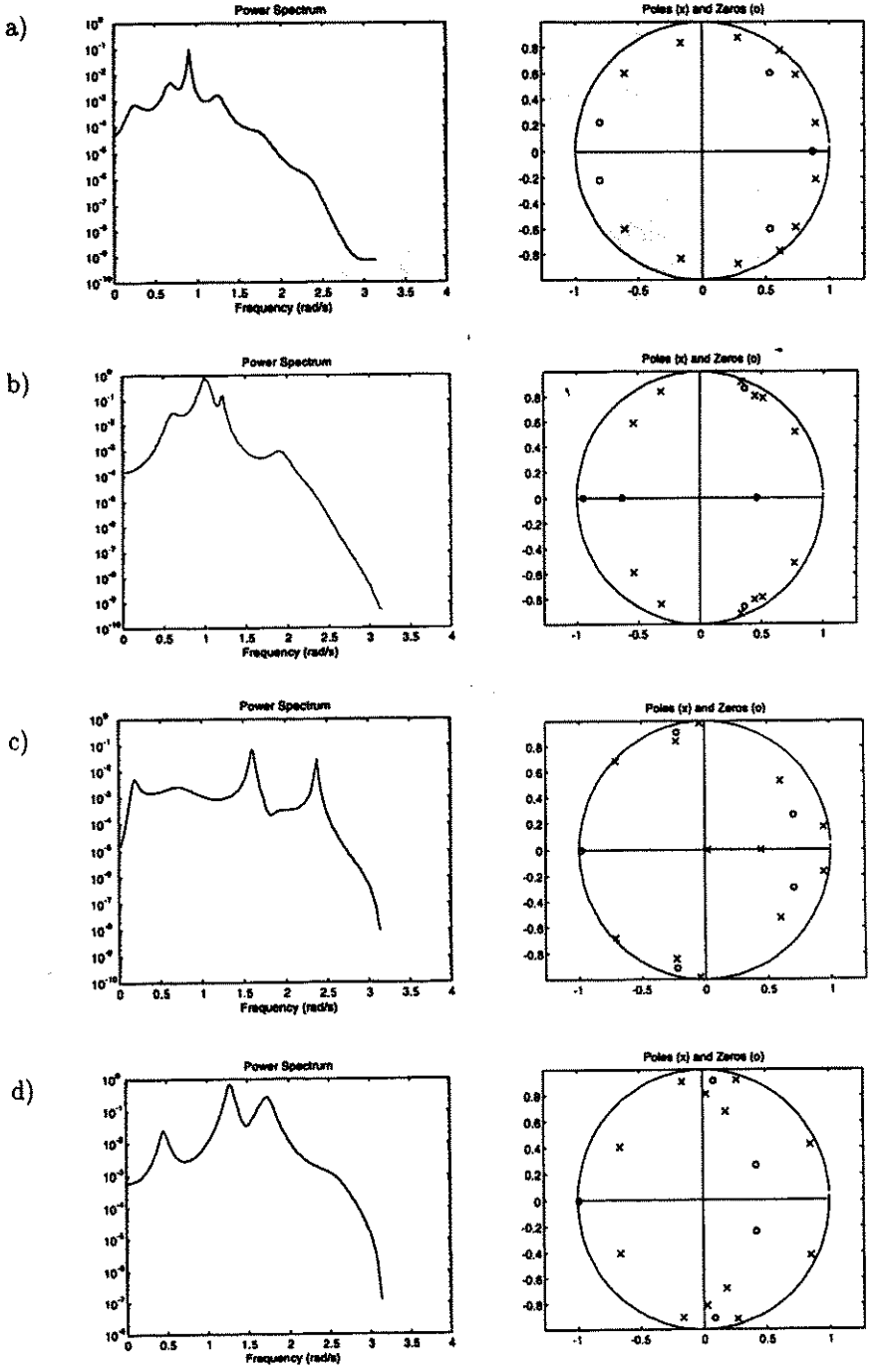


Fig. 3. Vocal tract transfer functions and pole-zero plots for: /n/ in /na/ (a), /a/ in /na/ (b), /m/ in /me/ (c), /e/ in /me/ (d), as obtained by a (12,5)-order ARMA model using LSMYWE.

[4, 20]. Fundamental frequency is set at about double (1 octave higher) than the mean F0 estimated from the esophageal speech signal. The inclusion of jitter and/or shimmer perturbations on the synthetic glottal signal may significantly improve the quality of produced speech.

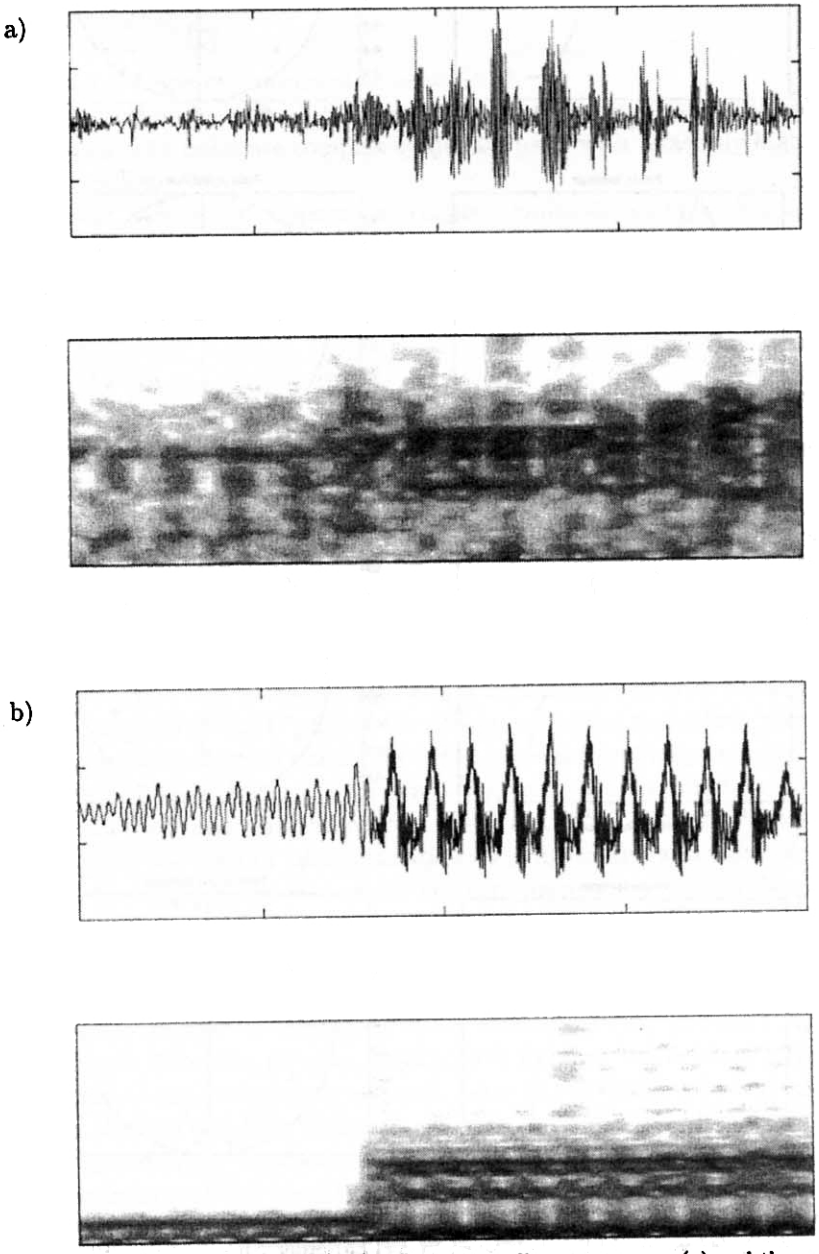


Fig. 4. Esophageal speech utterance /na/ and corresponding spectogram (a) and the resynthesized signal and its corresponding spectogram (b).

The synthesized speech signal is computed in overlapping frames using triangular weighting of the current and next frame's analysis window. Figure 4 displays an original esophageal /na/ utterance and the resynthesized one together with their corresponding spectograms. As it can be observed, the formantic structure is preserved, and a higher and regular F0 pattern appears after the resynthesis procedure.

Early results of subjective tests on the synthetic speech produced with the proposed method show that great improvement in voice quality and spoken utterances' intelligibility is achieved, including voiced consonantal sounds. Also, the resynthesized speech preserves the cues needed for speaker identification. The LSMYWE method shows sufficient ability in tracking both poles and zeroes of the vocal tract acoustic filter. A technique which uses more generalized speech production models such as ARMAX and Box–Jenkins models together with smoothing formant and antiformant trajectories prior to convolution with the synthetic glottal signal is currently investigated.

5. Conclusions and further work

This work focused on the investigation and methods for extraction of major features for Greek esophageal speech.

Fundamental frequencies were found at about 70 Hz confirming the results of previous works arguing that esophageal speech sounds about 1 octave lower than normal speech. The Center-clipped Autocorrelation and the "Hilbert–Envelope" methods were found to be more efficient among various well-known methods of F0 extraction.

Slopes of power envelope vs. time during phonations served as a measure of voice dynamics and training progress. Mean values were estimated at about $-90\,\mathrm{dB/s}$ exhibiting a progressive increase from the less to the more trained speakers.

Formant profiles were studied for all Greek vowels and results showed a general coincidence with values of normal speech and verified the general distinctive features of Greek vowels to other languages' ones. The facts enforce the use of formant extraction method for esophageal speech recognition.

A new method for esophageal speech enhancement which is based on resynthesizing voiced segments using ARMA modeling of vocal tract and a higher pitch glottal waveform signal was also proposed and exhibited encouraging results.

Concurrent work includes study of major esophageal voice features such as jitter, shimmer, S/N ratios, intonational characteristics, e.t.c. under various conditions of the patient's surgical treatment procedure, period of training, e.t.c.

The proposed method for voice quality enhancement is revised under the use of extended ARMAX and Box–Jenkins models and variations of formant trajectories' extraction methods and glottal waveform signals. Accordingly, subjective quality tests are going to take place for the assessment of resynthesized speech. Further, the method is intended to be implemented on a DSP for real-time use.

References

[1] N. Bi, Y. Qi, *Alaryngeal speech enhancement based on spectral substitution*, ASA 127th Meeting M.I.T. 1994 June 6-10.

[2] N. Bi, Y. Qi, *Application of speech conversion to alaryngeal speech enhancement*, IEEE Trans. Speech & Audio Processing, **5**, 2, 97–105 (1997).

[3] Y.M. Cheng, D. O'Shaughnessy, *Automatic and reliable estimation of glottal closure instant and period*, IEEE Trans. AS SP, **37**, 12, 1805–1815, December 1989.

[4] P. Cook, *Identification of Control Parameters in an Articulatory vocal tract model with applications to the synthesis of singing*, Ph.D. Thesis, Stanford Univ., 1991.

[5] J. Flanagan, *Speech analysis, synthesis and perception*, Springer Verlag, 1972.

[6] S. Furui, *Digital speech processing, synthesis and recognition*, Marcel Dekker, Inc., 1989.

[7] S. Furui, M. Sondhi Mohan, *Advances in speech signal processing*, Marcel Dekker, Inc., 1992.

[8] A. Iivonen, *Articulatory vowel gesture presented in a psychoacoustical F1/F2-space*, Studies in Logopedics and Phonetics, Univ. Of Helsinki, **3**, 19–44 (1992).

[9] H. Jarkin, M. Galler, N. Niedzielski, *Enhancement of esophageal speech by injection noise rejection*, Proc. ICASSP'97, Munich, 1997.

[10] S. Kay, *Modern spectral estimation*, Prentice-Hall Signal Processing Series, 1988.

[11] A. Leinonen, *Intonational patterns and voice quality in esophageal speech*, Studies in Logopedics and Phonetics, Univ. of Helsinki, **3**, 151–159 (1992).

[12] K. Mourikis, *Phonetic rehabilitation of alaryngeal people*, 4th European Interuniversity Symposium of "Head and Neck Cancer: Improvements of Logoregional Control", Thessaloniki 1994.

[13] A. Oppenheim, R. Schafer, *Discrete time signal processing*, Prentice-Hall International 1989.

[14] C. Pastiadis, G. Papanikolaou, *A preliminary study on Greek esophageal speech and a method for voice quality enhancement*, AES 102nd Convention Preprint, Munich, March 1997.

[15] Y. Qi, R. Fox, *Analysis of nasal consonants using perceptual linear prediction*, Journal of the Acoustical Society of America, **91**, 3, 1718–1726 (1992).

[16] Y. Qi, *Replacing tracheoesophageal voicing sources using LPC synthesis*, Journal of the Acoustical Society of America, **88**, 3, 1228–1235 (1990).

[17] Y. Qi, B. Weinberg, N. Bi, *Enhancement of female esophageal and tracheoesophageal speech*, Journal of the Acoustical Society of America, **98**, 5, 2461–2465, 1995.

[18] Y. Qi, B. Weinberg, *Characteristics of voicing source waveforms produced by esophageal and tracheoesophageal speakers*, Journal of Speech & Hearing Research, **38**, 536–548 (1973).

[19] J. Robbins, H. Fisher, F. Blom, M. Singer, *A comparative study of normal, esophageal and tracheoesophageal speech production*, Journal of Speech and Hearing Disorders, **49**, 202–210 (1984).

[20] P. Rubin, L. Goldstein, *Articulatory synthesis (ASY)*, Haskins Laboratories.

[21] M. Mohan Sondhi, *New methods of pitch extraction*, IEEE Trans. Audio and Electroacoustics, **16**, 16, 262–266 (1968).

[22] I. Titze, H. Liang, *Comparison of F0 extraction methods for high-precision voice perturbation measurements*, Journal of Speech and Hearing Research, **36**, 1120–1133 (1993).

[23] R. Tull, J. Rutledge, J. Mahler, *Female alaryngeal speech enhancement for improved speaker identification using linear predictive synthesis*, ASA 129th Meeting Washington D.C. 1995 May 30–June 6.

[24] B. Weinberg, Y. Horii, B. Smith, *Long-time spectral and intensity characteristics of esophageal speech*, Journal of the Acoustical Society of America, **67**, 5, 1781–1784 (1980).