A STUDY OF TEMPORAL STRUCTURE OF GLOTTAL FLOW DERIVATIVE ESTIMATES OBTAINED VIA INVERSE FILTERING

Emir Turajlic

Sarajevo School of Science and Technology,
Bistrik 7, 71000 Sarajevo, Bosnia and Herzegovina
phone: + (387) 33563030, fax: + (387) 33563033,
emir_turajlic@hotmail.com

Saeed Vaseghi

Department of Electronics and Computer Engineering,
Brunel University Uxbridge, UB8 3PH, United Kingdom
phone: + (44) 1895 274 000, fax: + (44) 1895 232806,
saeed.vaseghi@brunel.ac.uk

ABSTRACT

This paper presents a comparative study of the temporal structure of the glottal flow derivative estimates in relation to an idealized view of voice source realizations as defined by Liljencrants-Fant's model. Specifically, we endeavor to ascertain the extent by which Liljencrants-Fant's model can be used to represent the glottal flow derivative estimates obtained via closed-phase pitch synchronous inverse filtering of recorded speech. The study includes several phonation types and two examples of voice pathology. The study has established the following. Due to the limited degrees of freedom, Liljencrants-Fant's model is only capable of adequately representing the "coarse" glottal pulse structure. The "fine" structural elements can constitute a considerable part of a glottal flow derivative realization, and we have presented evidence that they contain information related to voice individuality. In addition, we have shown that LF-parameters do not always accurately portray significant events in the vocal fold dynamics.

1. INTRODUCTION

The estimation and modeling of glottal excitation waveform are some of the most important and challenging areas of speech processing. Although the larynx is not easily accessible, a range of techniques have been developed to enable the study of vocal fold dynamics, e.g. electroglottography [15], electromagnetic-glottography [15], transillumination and high-speed imaging [9].

The inverse filtering of recorded speech is undoubtedly the most popular approach for estimation of the glottal excitation signals. It is a non-invasive method that does not require bulky or expensive equipment. However, inverse filtering implies strong assumptions regarding the speech production system. Correspondingly, the inverse filtering results are regarded as the glottal excitation estimates so as to differentiate them from the actual glottal flow derivative waveforms. This might sound somewhat trivial, but the distinction is significant in the context of two important applications, reconstruction of the glottal excitation signal and glottal flow derivative parameterization. Essentially, the inadequacies in the source–filter model of speech production and the inaccuracies in the implementation of inverse filter model (e.g. estimation of the closed-phase intervals) are largely manifested in the waveforms of voice source estimates. In voice quality profiling, whereby one seeks to obtain a parametric representation for the perceived voice textures, the vocal tract artifacts and the artifacts of non-linear time-varying source-filter coupling are seen as degradations that corrupt the actual voice source signal and conceal the true voice source parameters. Conversely, in speech synthesis, glottal excitation estimates represent the actual signals that need to be adequately modeled in order to achieve a faithful speech reconstruction.

With regards to the temporal structure of voice source waveform, Zannger *et al*. have shown that "distorted" sounds in singing voice are associated with complex voice production mechanisms that tend to produce structurally "rich" glottal flow realizations [18]. Švec *et al*. have demonstrated, by means of videokymography, that even healthy speakers commonly exhibit significant deviations from the idealized vocal fold behavior [14]. Distinctly adducted phonations often carry complex vibratory patterns including manifestations of vocal fold "ripples". *Creaky* phonations can have irregular vocal fold vibrations and are often linked with sub-harmonic patterns. Their findings also show that healthy larynges are rarely symmetric and that the phase delay between the vocal folds can have significant influence on glottal flow dynamics and voice texture. In fact, in extreme cases the voice can sound completely hoarse.

The rest of this paper is organized as follows. Section 2 describes Liljencrants-Fant's glottal flow derivative model. Brief overviews of closed-phase pitch-synchronous inverse filtering and formant modulation analysis are presented in sections 3 and 4, respectively. In Section 5, both techniques are employed on a range of voice qualities to enable a qualitative evaluation of the temporal structure of the voice source estimates. Section 6 concludes the paper.

2. LILJENCRANTS-FANT GLOTTAL MODEL

Over the past decades of research in glottal models, a number of solutions have been proposed, [6], [10], [12], [16]. Liljencrants-Fant's (LF) model [6] is the most widely adopted voice source representation, and thus it is selected as a subject of our study. Liljencrants-Fant's model is defined as:

$$\begin{aligned} v(t) &= E_0\, e^{\alpha t} \sin(\omega_g\, t) && ,0 \le t < T_e \\ v(t) &= \frac{Ee}{\varepsilon T_a}\left[e^{-\varepsilon (t-T_e)} - e^{-\varepsilon (T_c - T_e)}\right] && ,T_e \le t < T_c \\ v(t) &= 0 && ,T_c \le t < T_0 \end{aligned} \qquad (1)$$

where the time origin, t=0 corresponds to the vocal fold opening onset. E_e and T_0 denote the maximum glottal airflow declination rate and the glottal cycle duration, respectively. The shape of the glottal flow derivative waveform is

determined by the *timing parameters*, T_p, T_e, T_c and T_a. The reminding two parameters are defined as $\omega_g = \pi / T_p$ and $\varepsilon = (1 - e^{-\varepsilon(T_c - T_e)}) / T_a$.

3. CLOSED-PHASE INVERSE FILTERING

According to the source–filter theory, the transfer function of the voiced speech can be expressed as:

$$S(z) = A\,G(z)\,V(z)\,R(Z) = \frac{A\,G(z)\,R(z)}{H(z)} \qquad (2)$$

where $G(z)$ denotes the z-transform of the glottal flow over a pitch period; A is the gain factor; $V(z) = 1/H(z)$ describes the minimum-phase all-pole vocal tract transfer function; $R(z)$ corresponds to the radiation load. The combined effect of glottal flow, radiation load and gain can be expressed as $b(n) = A\,g(n) * r(n)$. Since radiation load can be represented by a differencing filter, $r(n) = \delta(n) - \delta(n-1)$ [8], [11], the sequence $b(n)$ describes a scaled glottal flow derivative over one pitch period. Thus, the *voiced* speech signal, $s(n)$ can be modeled as:

$$s(n) = \sum_{k=1}^{p} h(k)s(n-k) + b(n) * \sum_{m=-\infty}^{\infty} \delta(n - mP) \qquad (3)$$

where P denotes duration in-between successive excitation impulses. During the glottal closed-phase, eqn. (3) is not driven by $b(n)$. Hence, glottal flow derivative can be estimated by inverse filtering the speech waveform with an all pole vocal tract model that is derived over the closed-phase interval [5], [17]. This is the main principle behind the closed-phase pitch-synchronous inverse filtering.

4. FORMANT MODULATION ANALYSIS

Formant modulation analysis is a term that describes the study of formant frequency movement within a glottal cycle. Since formant modulation (movement) is a result of time-varying non-linear source/vocal tract coupling, it is expected to be more prominent during the glottal *open-phase* than during the *closed-phase*, when the vocal folds are closed [3]. Correspondingly, the closed-phase of a glottal cycle can be estimated as a region during which formants are relatively "stationary". In addition, the extent of formant modulation during the glottal open phase can be used to indicate the level of source/filter coupling during speech production.

Here, we present a brief overview of formant modulation analysis. A more detailed discussion is given in [13]. Formant modulation analysis is performed on the 1st resonant frequency of vocal tract. Compared to other resonances, it exhibits the strongest dynamics after the onset of the open phase and attains the highest degree of stationarity during the closed-phase [13]. The 1st formant trajectory is estimated over a glottal cycle duration using a one-sample-shift sliding covariance-based linear prediction analysis. The analysis is initiated at one sample after an identified GCI mark, and is continued until the analysis window reaches the next GCI mark. Hence, there are N-N_w number of windows over each glottal cycle, where N and N_w denote the pitch period and the analysis window length, respectively. N_w is set to $N/4$. Vocal tract coefficients are estimated for each analysis window using an all-pole vocal tract model of order, p=14. Subsequently, the 1st formant trajectory is estimated by performing a Viterbi search on a space constrained to the four lowest poles with bandwidths less than 500 Hz;

Estimation of glottal closed-phase intervals is based on a statistical analysis of the 1st formant trajectory values. The first step in this statistical approach is to identify a region of the formant trajectory that exhibits the highest degree of local stationarity. This region of the glottal cycle is referred to as the *initial stationary formant region* (ISFR) and it is estimated via the following algorithm:

$$ISFR = \arg\min_{n} \sum_{i=n}^{n+4} \left| F(i) - F(i-1) \right| ,\ 1 \le n < N - N_w - 5 \qquad (4)$$

where $F(i)$ denotes the 1st formant's value at the i^{th} sample after the instant of glottal closure. A conservative amount of data (five formant values) is used in an attempt to avoid taking formant values that are possibly outside the stationary region. Subsequently, a statistical model of formant modulation is developed for the *initial stationary* region. Gaussian distribution is used for this purpose. In the next stage, the *initial stationary* region is expanded with the neighboring points that are statistically similar to the *initial stationary* region. The expansion is done by a one-sample-shift using the following principle: if the next formant value is less than two standard deviations away from the mean value of the statistical model, it is associated with the stationary region. As the *initial stationary* region is expanded to the right, the statistical model of the stationary formant region is adapted to include the "new points". The instant where the formant deviation from the statistical mean exceeds the threshold value is used to mark the end of the glottal closed-phase interval. Subsequently, the stationary formant region is expanded to the left to identify the closed-phase onset. However, this time the statistical model is not adapted as it is already well established.

5 EXPERIMENT AND RESULTS

The closed-phase pitch synchronous inverse filtering and the formant modulation analysis are performed on a segment of sustained vowel /a/ for 5 male speakers. The dataset includes *modal* voice, *creaky* voice, *breathy* voice, and two examples of voice disorder, *laryngeal cancer* and *vocal fold paralysis*, sampled at F_s =10 kHz. All data recordings were performed in a professional single-wall sound room using a Bruel and Kjaer model 4113 microphone located 6 inches from the speaker's lips. Vocal tract poles are estimated over the closed-phase regions, as determined by formant modulation analysis, using the covariance method of linear prediction with a 14th order predictor. Although we have considered other inverse filtering methods, such as PSIAIF method [1], and other linear prediction orders for the vocal tract model, the results of our tests show that the selected technique yields the best performance. The results of formant modulation analysis and inverse filtering are presented in Figures 1-5. Note that the formant trajectory graphs correspond to 75% of glottal cycle duration, as prescribed by formant modulation analysis procedure. The time domain labeling of both panels is referenced to the identified glottal closure instant.

In each of five examples, the estimated formant trajectories exhibit clearly defined formant stationary regions. The most extensive formant modulation is observed in *vocal fold paralysis* and *breathy* voice examples. On the other end

Table 1 - Liljencrants-Fant's describing the synthetic waveforms in Figure 6, expressed as a percentage of glottal cycle duration

Stimuli	T_p [%]	T_e [%]	T_c [%]	T_a [%]	F_0 [Hz]
Creaky voice	8.17	9.15	34.13	13.7	76.34
Breathy voice	64.22	78.90	98.00	11.0	91.74
Vocal fold paralysis	43.00	61.00	86.00	15.0	103.09
Laryngeal cancer	48.00	75.00	98.33	10.2	169.50
Modal voice	36.26	40.66	59.34	9.80	111.11

Table 2 - Modeling SNR for six speakers

Creaky	Breathy	V.f. paralysis	Cancer	Modal
4.10 dB	8.92 dB	14.49 dB	6.88 dB	9.86 dB

of the scale is the *modal* voice with the weakest formant modulation. An interesting observation is that the stationary formant regions do not always coincide with the closed-phase intervals according to the Liljencrants-Fant's representation of glottal flow derivative waveforms. In the instances of *laryngeal cancer* and *breathy* voice, the stationary formant regions extend well beyond the nominal closed-phases, whereas for *modal* and *creaky* voices, the formant modulation onsets occur prior to the nominal open phases. In the *breathy* voice example, inspection of the glottal flow derivative estimate suggests that vocal folds do not fully close. However, the results of formant modulation analysis reveal a clearly distinct region in which formants are stationary indicating a lack of source-filter coupling and a complete vocal fold closure. In both, *modal* and *creaky* voices, the formant modulation onset occurs before the onset of the nominal open phase and coincides with the onset of formant ripple. Thus, we are lead to infer that the vocal folds must have been partly open during the nominal closed-phases. In *laryngeal cancer* and *creaky* voices, a high degree of turbulence is present in the voice source estimates suggesting a narrow and parallel vocal fold opening, rather than a triangular opening. The *laryngeal cancer* also exhibits a specific phenomenon that is not found in any other speaker; the voice source signal is highly irregular and two distinct types of glottal flow derivative realizations can be observed. Ultimately, the reason for the varied displays of the glottal flow derivative waveforms relates to the fact that the laryngeal settings, geometry, and physiology are different for each individual [3], [7].

We have employed a signal to noise ratio measure to establish the extent by which the Liljencrants-Fant's model can be used to represent the voice source estimates. Firstly, the glottal flow derivative estimates are parameterized using Alku, and Vilkman's *direct estimation* method [2]. Manual corrections were made when deemed necessary. The results of parameterization are displayed in Table 1. Subsequently, the Liljencrants-Fant's waveforms are subtracted from the glottal flow derivative estimates to obtain the modeling residual signals, i.e. $v_r(n) = v_g(n) - v_{LF}(n)$. The modeling SNR values are obtained via (5) and presented in Tables 2.

$$SNR = 10\log_{10}\left(\sum_{i=0}^{N-1} v_g^2(n) \Big/ \sum_{i=0}^{N-1}\left(v_g(n) - v_{LF}(n)\right)^2\right) \qquad (5)$$

,where N refers to the glottal cycle length.

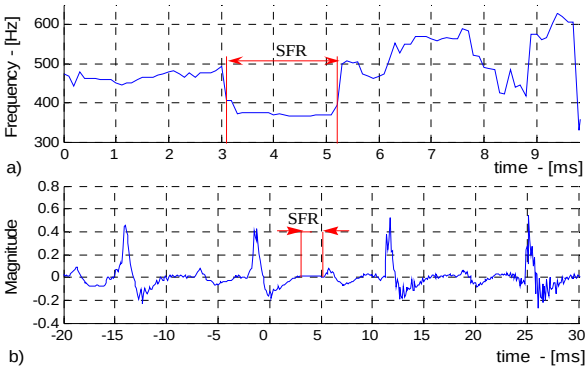

Figure 1 - *Creaky voice; a) 1st formant b) glottal flow derivative*

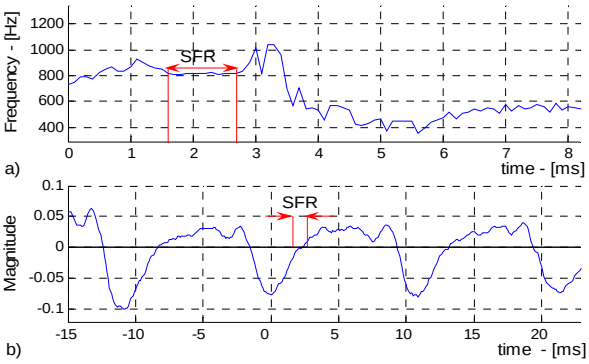

Figure 2 - Breathy voice; *a) 1st formant; b) glottal flow derivative*

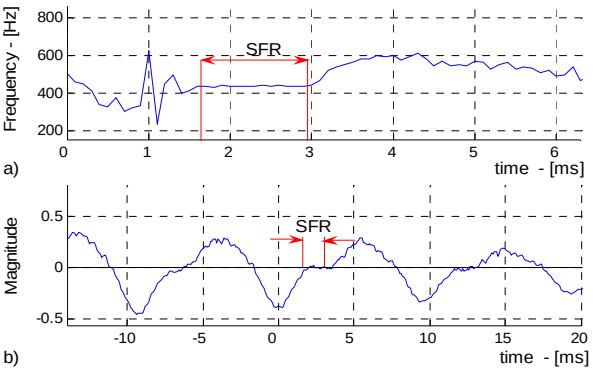

Figure 3 - V.f. paralysis; *a) 1st formant; b) glottal flow derivative*

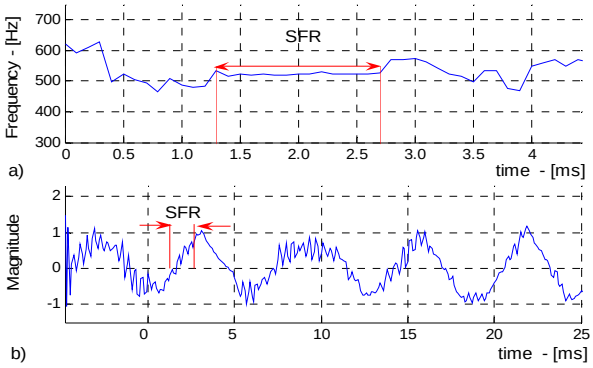

Figure 4 - Laryn. cancer; *a) 1st formant b) glottal flow derivative*

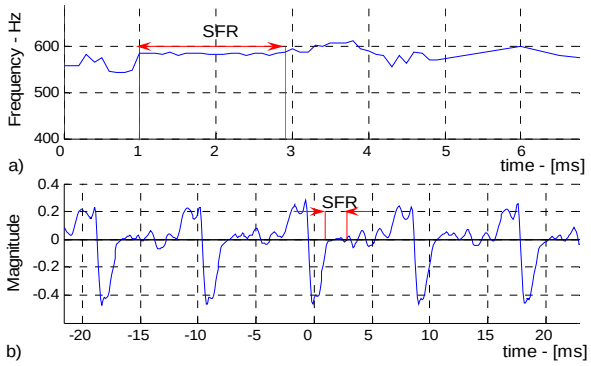

Figure 5 - Modal voice; *a) 1st formant; b) glottal flow derivative*

In Figure 6, we have displayed the estimated glottal flow derivative waveform, the synthesized Liljencrants-Fant's waveform, and the corresponding LF modeling residue, for each of 5 speakers. Note that T_{f0} denotes the glottal opening instant obtained via formant modulation analysis, while T_e marks the glottal closure instant. Since there is more than 10 dB difference between the best (*vocal fold paralysis*) and the worst (creaky voice) modeled voice source signal, we are inclined to suggest that the ability of the Liljencrants-Fant's model to represent the voice source signal may be speaker dependant. In order to substantiate this proposition, a study of LF residue waveforms needs to be conducted, as in [13]. In [13], the authors have focused on the *modal* voices, only. Their findings indicate that the first formant ripple and aspiration noise are the predominant features of the LF residual waveforms. Here, the term "ripple", also known as the first formant ripple, describes a sinusoidal-like perturbation in the glottal derivative waveform due to the time-varying non-linear coupling of the glottal flow and the vocal tract [3], [4]. Given that our study includes a wider range of voice quality types, we are able to conduct a more conclusive analysis of the residue waveforms.

Interestingly, in relation to *modal* voice, our results are in accord with those presented in [13]. However, in other examples, we have also identified the inadequacy of LF model to represent the complex temporal features in the voice source signal as jet another significant contributor to modeling error. In *breathy* voice, the ripple frequency is not anywhere near the formant frequencies as it is the case with *modal* voice. The graph in Figure 2*a*) shows that there is very little formant modulation during glottal open phase. Thus, we believe that the observed "ripple" constitutes an integral part of the speaker's voice source signal. *Creaky* voice is an interesting case as well. It contains comparable amounts of first formant ripple, modeling error* and aspiration noise. The first formant ripple dominates over more than the first half of the opening phase, while the modeling error and aspiration noise occupy the regions just prior and after the glottal closure instant, respectively. All three residual elements are clearly visible and seem to exist in temporal isolation. In the *laryngeal cancer* instance, the modeling residue waveform exhibits a high degree of irregularity. The formant ripple is a dominant

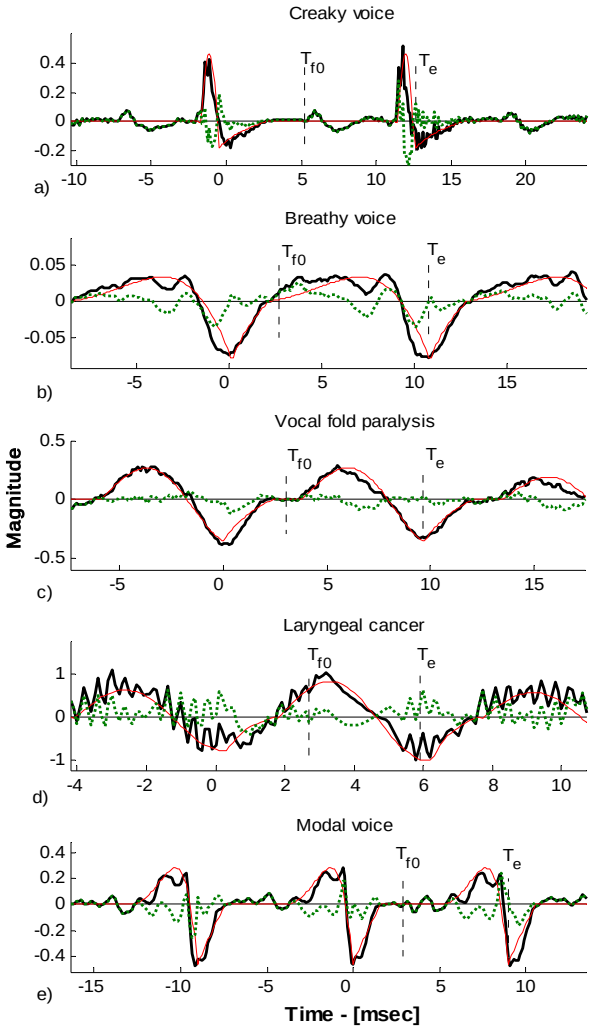

Figure 6 - *Glottal flow derivative estimates* (solid thick line), *synthesized LF waveforms* (solid thin line), *and the corresponding LF modeling residue* (dotted line).

residue feature only for the middle of the three glottal pulses. In other pulses, high frequency aspiration noise and modeling error are the principal elements of the residual structure. The last remaining subject of our analysis, *vocal fold paralysis*, displays by far the most idealistic voice source waveform. In addition, its modeling residue does not contain any significant amounts of formant modulation artifacts nor turbulent components related aspiration noise. Thus, the modeling SNR is notably higher than in other examples. Overall, these results show that the main residual features, namely, formant ripple, aspiration noise and modeling error, are a direct consequence of an over simplistic view of vocal fold realization that is adopted by the Liljencrants-Fant's model. The relative energy distribution of the individual residual elements exhibits drastic variation across speakers and phonation types. Even though a number of study signals considered in this paper is relatively small, we deem that this fact alone constitutes a notable evidence that the fine glottal flow derivative structure (LF

* For the purpose of simplicity, the term, *modeling error* is from here on used to specifically denote those features of the residual signal that can not be attributed to either aspiration noise or the formant ripple.

residue) might convey important information related to speaker individuality and possibly voice quality.

6. CONCLUSION

In this paper, we have taken a critical view against the "de facto" model of glottal excitation, i.e. Liljencrants-Fant's model. Closed-phase pitch-synchronous inverse filtering and a formant modulation analysis technique are employed on a range of voice qualities types, including two examples of laryngeal pathology, to enable a qualitative evaluation of the temporal structure of glottal excitation estimates. The results of our study suggest that due to the inherent complexity of glottal flow realizations, the inadequacies of the source–filter model of speech production and the inaccuracies in the implementation of inverse filtering, more often than not, voice source estimates do not completely comply with the idealized waveforms of Liljencrants-Fant's glottal flow derivative model. In the best of circumstances, Liljencrants-Fant's model provides enough degrees of freedom to adequately represent only the general shape or the "coarse structure" of the glottal flow derivative waveforms. The fact that Liljencrants-Fant's model can not represent complex voice source realizations nor the formant modulation ripples is a serious deficiency of this model. In the LF representation, the fine glottal flow derivative structure is discarded and correspondingly, some of the information related to the voice individuality and voice quality is inevitably lost. Furthermore, formant modulation analysis has shown that Liljencrants-Fant's parameters do not always accurately identify the significant events in the vocal fold dynamics, and thus, the process of LF-based voice source parameterization carries an inherent degree of fallibility. Presumably, these limitations are manifested in the qualities of LF-based speech synthesis and related voice quality conversion methods. Thus, we deem that a more sophisticated model is required to satisfy the requirements of the state of the art speech processing applications.

7. REFERENCES

[1] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering", *EUROSPEECH*, 1081-1084, 1991.

[2] P. Alku, and E. Vilkman, "Effects of bandwidth on glottal airflow waveforms estimated by inverse filtering," *J. Acoust. Soc. Am*. 98, 763-767, 1995.

[3] T.V. Ananthapadmanabha and G. Fant, "Calculations of true glottal flow and its components", *Speech Comm*., 1:167-184, 1982.

[4] D.G. Childers, and C. Wong, "Measuring and Modeling Vocal Tract Interaction", *IEEE Transactions on Biomedical Engineering*, 41, 663-671, 1994.

[5] K.E Cummings and M.A. Clements, "Analysis of Glottal Waveforms Across Stress Styles", *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 369-372, Albuquerque, 1990.

[6] G. Fant, J. Liljencrants and Q. Lin, "A four-parameter model of glottal flow", *STL-QPSR 4*, pp. 1-13, 1985.

[7] G. Fant and Q. Lin, "Glottal source–vocal tract acoustic interaction", STL-QPSR, (1):13-27, 1987.

[8]J.L Flanagan, "Speech Analysis, Synthesis and Perception", 3ed ed., New York: *Springer Verlag*, 1972.

[9] S. Kiritani, H. Imagawa, and H. Hirose, "Vocal cord vibrations and voice source characteristics – observations by a high speed digital recording", Proceedings of the International conf. on Spoken Language Processing, pp.61-64, Japan, 1990.

[10] D.H. Klatt and L.C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *J. Acoust. Soc. Am. 87(2)*, pp. 820-57, 1990.

[11] Q.G. Lin, "Speech Production Theory and Articulatory Speech Synthesis", *PhD dissertation*, Royal Institute of Technology, Stockholm, Sweden, 1990.

[12] P.H. Milenkovic, "Voice source model for continuous control of pitch period", *J. Acoust. Soc. Am.* 93(2), pp. 1087-96, 1993.

[13] M. D. Plumpe, T.F. Quartiery, D.A. Reynolds, "Modeling of the Glottal Flow Derivative Waveform with Applications to Speaker identification", *IEEE Transactions on Speech and Audio Processing*, Vol. 7,no. 5, 1999.

[14] J.G. Švec, H.K, Schutte, F. Šram, "On Vibration Properties of Human Vocal Folds: Voice Registers, Bifurcations, Resonance Characteristics, Development and Application of Videokymography", *PhD thesis*, University of Groningen, Netherlands [ISBN: 90-367-1235-1], Ch. 8, pp. 91-93, 2000.

[15] I.R., Titze, B.H. Story, G.C. Burnett, J.F. Holzirichter, L.C Ng, and W.A. Lea, "Comparison between Electroglottography and electromagnetic glottography", *J. Acoust. Soc. Am*, vol. 107, no. 1, pp. 581-588, 2000.

[16] R.N.J. Veldhuis, "A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation", *J. Acoust. Soc. Am*. 103, pp. 566-71, 1998.

[17] D.Y. Wong, J.D. Markel and A.H. Gray Jr., "Least-Squares Glottal Inverse Filtering from the acoustic Waveform", *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-27, no.4, pp. 350-355, Aug. 1979.

[18] B.D. Zangger, J. Sundberg, P.A. Lindestad, M. Thalen, "Vocal fold vibration and voice source aperiodicity in phonatorily distorted singing", *Speech, Music and Hearing*, TMH-QPSR, vol. 45: 87-91, 2003.