ESTIMATION OF HARMONIC AND NOISE COMPONENTS OF THE GLOTTAL EXCITATION

R. Sousa[1], A. Ferreira[1] and P. Alku [2]
[1]University of Porto – School of Engineering, Porto, Portugal
[2] Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland

Abstract: **This paper describes an algorithm which enables harmonic and noise splitting of the glottal excitation of voiced speech. The algorithm utilizes a straightforward harmonic and noise splitter which is utilized prior to glottal inverse filtering. The results show improved estimates of the glottal excitation in comparison to a known inverse filtering method.**

Keywords: **Voice quality, voice diagnosis, glottal inverse filtering**

I. INTRODUCTION

Since the glottal volume velocity waveform serves as the source of (voiced) speech, it has an essential role in the production of several acoustical phenomena such as the regulation of vocal intensity [1], voice quality [2], the production of different vocal emotions [3] and voice pathologies detection related to vocal fold changes [4]. Therefore, accurate analysis and parameterization of the glottal pulseform is beneficial in several areas of speech science including both healthy and disordered voices. In this paper, two techniques are combined to yield an algorithm that estimates the harmonic and noise components of the glottal pulse. These techniques decompose the signal into a harmonic and noise component and gives rise to better glottal pulse estimations. This new algorithm was tested with synthetic and natural voices in order to characterize the algorithm behavior against an acoustic diversity.

II. METHODS

A. Algorithm overview

The main goal of the study is to develop an algorithm that splits the waveform of the estimated glottal airflow into a harmonic and a noise component. The block diagram of the method is shown in Fig. 1.

First (block 1), the speech pressure signal is divided into a harmonic and a noise component [5]. Secondly (block 2), the obtained harmonic component of the speech signal, denoted by h(n) in Fig. 1, is used as an input to glottal inverse filtering which yields an estimate of the vocal tract inverse filter (an FIR filter), denoted by V(z) in Fig. 1. Inverse filtering is computed with a previously developed automatic algorithm, Iterative Adaptive Inverse Filtering (IAIF) [6]. Thirdly, this FIR filter is used in order to cancel the effects of the vocal tract from three signals: both from the harmonic and noise components obtained from the harmonic-noise splitter, and from the original speech pressure waveform. By further canceling the lip radiation effect using an integrator whose transfer function is simply given by $H(z)=1/(1-0.99z^{-1})$, three glottal signals are obtained: the glottal pulse harmonic component, the glottal pulse noise component, and the glottal pulse, which are denoted in Fig. 1 by $g_h(n)$, $g_r(n)$, and g(n), respectively. Equations (1) to (4) express the resulting signals in Fig. 1.

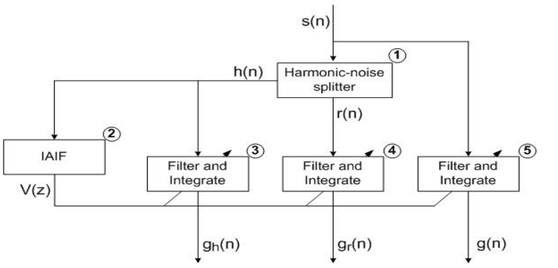

Fig. 1: Main block diagram of glottal harmonic-noise splitter. Signals s(n), h(n) and r(n) denote, respectively, the speech signal and its harmonic and noise components. Signals g(n), g_h(n) and g_r(n) denote, respectively, the glottal excitation, and its harmonic and noise components. V(z) denotes the vocal tract transfer function. IAIF denotes the glottal inverse filtering algorithm [6].

$$s(n) = h(n) + r(n) \tag{1}$$

$$g(n) = v(n) * \ell(n) * [h(n) + r(n)] \tag{2}$$

$$g(n) = v(n) * \ell(n) * h(n) + v(n) * \ell(n) * r(n) \tag{3}$$

$$g(n) = g_h(n) + g_r(n) \tag{4}$$

The parameters *v(n)* and ℓ(n) denote the impulse response of the inverse model of the vocal tract and lip radiation effect, respectively. Equation (1) represents the harmonic-noise model, which serves as the basis for the harmonic-noise splitter. Inverse filtering is represented by equation (2). Equations (3) and (4) show that the glottal excitation consists of harmonic and noise components.

The harmonic-noise splitter is based on a model of the harmonic structure of speech, which is parameterized in frequency, magnitude and phase [5]. The block diagram of the harmonic-noise splitter is depicted in Fig. 2.

In the first stage (block 1), the time domain input

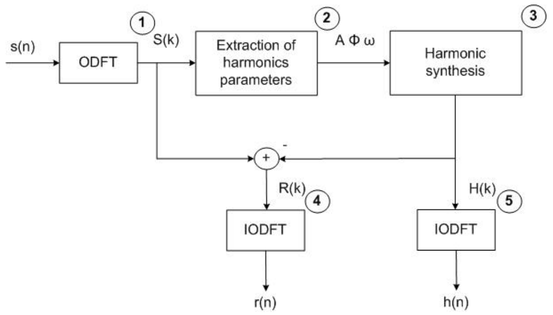

Fig. 2: Block diagram of the harmonic-noise splitter.

signal is transformed into the frequency domain using an Odd-Discrete Fourier Transform (ODFT) [7]. ODFT is obtained by shifting the frequency index of the Discrete Fourier Transform (DFT) by half a bin:

$$X_o(k)=\sum_{n=0}^{N-1} x(n)e^{-j\frac{2\pi}{N}(k+\frac{1}{2})n}, \quad k=0,1,...,N-1 \tag{5}$$

where the time-domain input signal is denoted by x(n) and the frame length is N. If x(n) is real, this frequency shift makes the DFT samples above π a perfect mirror (in the complex conjugate sense) of the DFT samples below π. A peak picking algorithm is used to estimate the harmonics of the ODFT amplitude spectrum. Next, the frequency, magnitude and phase of each harmonic are extracted (block 2) [7]. These parameters are then used to synthesize the spectrum of the harmonic structure of the input signal s(n) (block 3). The spectrum of each individual sinusoid is synthesized using the parameters extracted from that harmonic.

The synthesized harmonic structure is subtracted from the signal s(n) and the result is regarded as the noise component. The spectra of both components are inverse transformed in order to get time-domain representations for the components (blocks 4 and 5).

B. Performance assessment

Experiments were conducted by using both synthetic and natural vowels. The estimated glottal excitation waveforms were parameterized with two known parameters: the Normalized Amplitude Quotient (NAQ) and the difference (in dB) between the amplitudes the first and second harmonic (DH12). The NAQ parameter is a time-based parameter that is extracted for each glottal pulse and it measures the pressedness of phonation from the ratio of the peak-to-peak flow and the negative peak amplitude of the flow derivative [8]. The DH12 parameter is a frequency domain quantity and it measures the decay of the voice source spectrum [9]. Both parameters are independent of time and amplitude shifts. The relative error was used for NAQ since this parameter is a time-domain quantity that is typically measured on the linear scale and the absolute error was used for DH12 because this parameter is typically expressed in the dB scale.

A synthesizer based on the source-filter and harmonic-noise models was used to generate a set of test vowels. The source generation was based on Liljencrants-Fant (LF) model [10]. The fundamental frequency F0 was varied from 100 Hz up to 400Hz with an increment of 10 Hz, in order to mimic both male and female speech. For each pitch, several vowel instances were generated by varying HNR from 9 dB up to 21 dB with an increment of 1 dB. The HNR is acquired as:

$$\mathrm{HNR} = 10\times\log_{10}\left(\frac{E_h}{E_r}\right) \tag{6}$$

E_h and E_r denote, respectively, the energy of the harmonic component and the noise component of synthetic speech. The values of the LF model were selected according to Gobl [11] in order to involve three different phonation types (breathy, normal and pressed). The vocal tract filter was adjusted to synthesize the vowel [a] (F1= 664 Hz, F2=1027 Hz, F3=2612 Hz). All the data were generated using the sampling frequency of 22.05 kHz.

In the second experiment, a database that included 39 sustained waveforms of the vowel [a] uttered by 13 subjects (7 males, 6 females) using breathy, normal and pressed phonation was used. The data were sampled with 22.050 kHz and a resolution of 16 bits. From these signals, the most stable segments with duration of 200 ms were selected for the voice source analysis.

III. Results

A. Experiments with synthetic voices

This section presents the results that were obtained for synthetic voices when the glottal source was estimated with IAIF and the proposed method. The NAQ error and DH12 error were determined separately for each phonation type. In order to compress the results, a set of ranges were defined for F0 and HNR and the individual values obtained inside these ranges were pooled together. For F0, the following three ranges were used: 100-200 Hz, 210-300 Hz, and 310-400 Hz. The first two ranges correspond to typical pitch used by males and females, respectively. The third range represents F0 values typical in voices produced by children. For HNR, the following three categories were used: 9-15 dB, 16-21 dB, and 22-27 dB. The first of these is typical for pathological voices while the second is characteristic to normal speech [12]. The last HNR range is related to voices which are highly periodic with a small amount of noise, such as the singing voice [13]. For each phonation type, the results are organized in tables that show the performance of NAQ or DH12 for the selected F0 and HNR ranges.

Tables 1 and 2 show that the proposed algorithm yields smaller DH12 errors for all the F0 and HNR combinations analysed from pressed vowels. The mean NAQ error was smaller with the proposed method also for all the F0 and HNR combinations except for three cases (F0 ranges 210-300 Hz and 310-400 Hz combined with HNR range of 16-21 dB; F0 range 310-400 Hz combined with HNR range 22-27 dB).

Table 1: NAQ mean relative error (in percentage) for IAIF and the proposed method in the analysis of pressed synthetic voices.

F0 (Hz)	IAIF HNR (dB)			Prop. Meth. HNR (dB)		
	9-15	16-21	22-27	9-15	16-21	22-27
100-200	27,8	14,8	22,6	13,0	11,2	15,5
210-300	52,8	27,5	75,6	21,2	38,4	60,4
310-400	64,7	68,9	131,3	55,9	101,1	151,0

Table 2: DH12 mean absolute error (in dB) for IAIF and the proposed method in the analysis of pressed synthetic voices.

F0 (Hz)	IAIF HNR (dB)			Prop. Meth. HNR (dB)		
	9-15	16-21	22-27	9-15	16-21	22-27
100-200	4,6	1,4	0,8	1,0	0,5	0,4
210-300	14,3	3,6	4,0	4,7	2,4	2,0
310-400	15,0	15,1	7,8	12,3	4,7	5,9

Tables 3 and 4 indicate that the proposed method yielded smaller errors for all the F0 and HNR ranges in the NAQ measurements in modal phonation.

Table 3: NAQ mean relative error (in percentage) for IAIF and the proposed method in the analysis of modal synthetic voices.

F0 (Hz)	IAIF HNR (dB)			Prop. Meth. HNR (dB)		
	9-15	16-21	22-27	9-15	16-21	22-27
100-200	38,2	21,3	9,3	14,2	8,0	4,7
210-300	68,9	38,2	16,7	24,4	11,4	10,8
310-400	68,5	54,5	36,5	38,3	24,0	28,0

Table 4: DH12 mean absolute error (in dB) for IAIF and the proposed method in the analysis of modal synthetic voices.

F0 (Hz)	IAIF HNR (dB)			Prop. Meth. HNR (dB)		
	9-15	16-21	22-27	9-15	16-21	22-27
100-200	7,2	0,9	0,8	1,6	1,4	0,7
210-300	15,7	6,4	3,8	5,4	1,0	1,9
310-400	9,4	16,3	11,9	16,9	4,0	2,9

For the DH12 error, the proposed method yielded larger distortion than IAIF only in two cases (F0 range of 100-200 Hz combined with the HNR range of 16-21 dB; F0 range of 310-400 Hz combined with HNR range of 9-15 dB).

Tables 5 and 6 show results from breathy voices that are in line with those observed for modal phonation: the mean NAQ error is smaller for the proposed method for all the F0 and HNR categories analysed and the mean DH12 error was also smaller with the proposed algorithm in comparison to IAIF for all the F0 and HNR combinations except for few cases (F0 range of 100-200 Hz combined with the HNR ranges of 16-21 dB and 22-27 dB; F0 range of 210-300 Hz combined with HNR range of 22-27 dB).

Table 5: NAQ mean relative error (in percentage) for IAIF and the proposed method in the analysis of breathy synthetic voices.

F0 (Hz)	IAIF HNR (dB)			Prop. Meth. HNR (dB)		
	9-15	16-21	22-27	9-15	16-21	22-27
100-200	56,9	37,0	16,5	25,8	11,6	12,0
210-300	77,9	68,2	23,9	46,9	17,9	13,3
310-400	83,8	80,7	45,8	54,4	31,6	18,9

Table 6: DH12 mean absolute error (in dB) for IAIF and the proposed method in the analysis of breathy synthetic voices.

F0(Hz)	IAIF HNR (dB)			Prop. Meth. HNR (dB)		
	9-15	16-21	22-27	9-15	16-21	22-27
100-200	9,8	4,6	2,4	5,3	5,1	4,3
210-300	32,8	24,3	4,5	15,7	6,7	5,5
310-400	21,0	28,2	13,3	20,8	9,1	5,7

In summary, the results obtained for the synthetic vowels show that the proposed method yields smaller mean NAQ and DH12 errors for the majority of the sounds analyzed. In particular, we highlight that the proposed method yields improved estimation accuracy in conditions with large amount of noise and for high-pitch voices. This accuracy improvement depends on the phonation type being more pronounced for modal voices.

B. Experiments with natural voices

Results computed from natural speech are shown in the form of time-domain waveforms by involving both the harmonic and the noise component yielded by the novel inverse filtering method.

Figures 3 and 4 show waveforms computed from utterances produced by a male and female speaker, respectively. From both of these figures one can observe that the harmonic component is smoother than the glottal excitation waveform. In addition, low frequency fluctuations are not present in the harmonic component and the noise component indicates amplitude perturbations at the instants of glottal closure.

IV. Discussion

Results obtained with synthetic voices show that the proposed method improves the estimation of the glottal waveform. The harmonic component given by the new algorithm is a more accurate estimate of the glottal source because the method is able to suppress the influence of noise which is always present in natural speech, particularly in pathological voices. The behavior of both algorithms was tested as a function of the noise level and fundamental frequency. The proposed method also

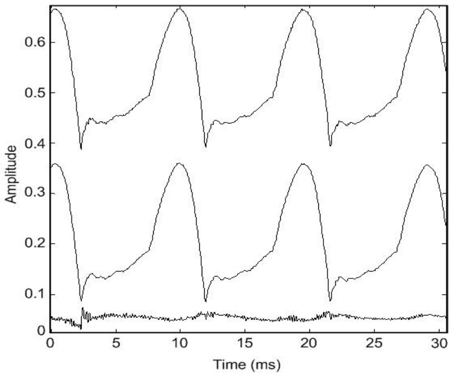

Fig. 3: Glottal excitation (top), its harmonic (middle) and noise (bottom) components estimated with the proposed method. A natural vowel [a] produced by a male speaker was used. The noise waveform is magnified 3 times for visual clarity.

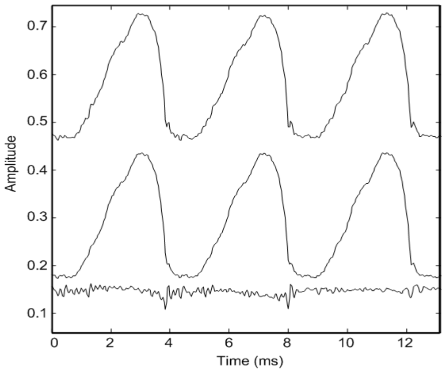

Fig. 4: Glottal excitation (top), its harmonic (middle) and noise (bottom) components estimated with the proposed method. A natural vowel [a] produced by a female speaker was used. The noise waveform is magnified 3 times for visual clarity.

enables joint estimation of the harmonic and noise components of the glottal waveform.

Drawbacks of the proposed method are due to the harmonic-noise splitter, which may pass noise to the harmonic component and itself is also sensitive to the noise level.

V. Conclusion

In this article, a method to estimate the glottal excitation based on a known automatic inverse filtering method, IAIF, and a harmonic-noise splitter was proposed. The new method was compared with IAIF in the estimation of the glottal excitation using experiments with both synthetic and natural vowels.

The proposed method enables joint estimation of the harmonic and noise components of the glottal waveform. These components may be used in the evaluation of pathological voices since the separation enables characterizing the vocal folds dynamics as a function of noise produced in the speech production process. In addition, the noise component estimated by the proposed method can be used in speech technology in order to improve the naturalness of synthetic speech.

Acknowledgements

This work has been developed in the context of a doctoral program supported by the Portuguese Foundation for Science and Technology under grant SFRH/BD/24811/2005 and project PTDC/SAU-BEB/104995/2008.

References

[1] F. Hodge, R. Colton and R. Kelley, “Vocal intensity characteristics in normal and elderly speakers,” *J. Voice*, vol.15, no.4, pp. 503–511, 2001.

[2] C. Gobl and A. Ni Chasaide, “The role of voice quality in communicating emotion, mood and attitude,” *Speech Commun.*, vol. 40, pp. 189–212, 2003.

[3] K. Scherer, “Vocal communication of emotion: a review of research paradigms,” *Speech Commun.*, vol. 40, pp. 227–256, 2003.

[4] P. Vilda, R. Baillo, V. Biarge, V. Luis, A. Marquina, L. Fernandez, R. Olalla and J. Llorente, “Glottal source biometrical signature for voice pathology detection,” *Speech Commun.*, vol. 51, pp.759–781, 2009.

[5] R. Sousa, “A new accurate method of harmonic-to-noise ratio extraction,” *Proc. of the Inter. Conf. on Bio-inspired Systems and Signal Proc.*, pp. 351-356, 2009.

[6] P. Alku, “Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering,” *Speech Commun.*, vol. 11, no. 2-3, pp. 109-118, 1992.

[7] A. Ferreira, “Accurate estimation in the ODFT domain of the frequency, phase and magnitude of stationary sinusoids,” *In: IEEE Workshop on App. of Signal Proc. to Audio and Acous.*, pp. 47–50, 2001.

[8] P. Alku, T. Bäckström and E. Vilkman, “Normalized amplitude quotient for parameterization of the glottal flow,” *J. Acoust. Soc. Am.*, vol. 112, no. 2, pp. 701-710, 2002.

[9] I. Titze and J. Sundberg, “Vocal intensity in speakers and singers,” *J. Acoust. Soc. Am.*,vol. 91, no. 5, pp. 2936–2946, 1992.

[10] G. Fant, J. Liljencrants, Q. Lin, “A four parameter model of glottal flow,” *STL-QPSR*, vol. 1, pp.1-13, 1985.

[11] C. Gobl, “A preliminary study of acoustic voice quality correlates,” *STL-QPSR*, pp. 9-21, 1989.

[12] J. Llorente, P. Vilda, F. Roldan, M. Velasco and R. Fraile, “Pathological likelihood index as a measurement of the degree of voice normality and perceived hoarseness,” *J. Voice*, vol.15, no. 7, pp. 1-11, 2009.

[13] J. Selby, H. Gilberta and J. Lerman, “Perceptual and acoustic evaluation of individuals with laryngopharyngeal reflux pre- and post-treatment,” *J. Voice*, vol. 17, no. 4, pp. 557–570, 2003.