GLOTTAL PULSE ESTIMATION - A FREQUENCY DOMAIN APPROACH

Sandra Dias, Aníbal Ferreira

Department of Electrical and Computer Engineering
University of Porto - Faculty of Engineering, Porto, Portugal

ABSTRACT

We describe a new frequency-domain glottal pulse estimation algorithm (FD-GPE) that takes advantage of the spectral diversity between sinusoidal and noise components of voice speech, that decouples the phase and magnitude contributions of both source and filter parts of voice production, and that relies on a hybrid Liljencrants-Fant / Rosenberg model of the glottal pulse. The FD-GPE algorithm is tested using synthetic and natural voiced vowels, and using two reference algorithms for comparison, IAIF and ZZT-CC. Results suggest that the performance of FD-GPE and IAIF is comparable, and that all algorithms give rise to glottal pulse estimates whose closing and closed phases are significantly different from those of idealized models. We conclude with next research steps.

1. INTRODUCTION

Locally stationary syllabic units of the speech, notably phonemes corresponding to vowels, can be modelled as the result of the convolution between an excitation acoustic signal and the impulse response resulting from the transfer function of the vocal tract filter. This model of voice production is often referred to in the literature as source-filter model [1], where the source represents the flow of the air leaving the lungs and passing through the glottis, and the filter represents the resonances of the vocal tract and lip/nostrils radiation.

Contrary to whisper, the glottal source in voiced vowels includes a periodic component consisting of glottal pulses which arise as a result of the vibration of the vocal folds when air flows through the glottis. The glottal pulses not only magnify the loudness of voice sounds, but also carry multidimensional information concerning notably the speaker identity [2] and its emotional state [3, 4], and the health condition of the vocal folds. This suggests that glottal pulse estimation from speech is very interesting in different areas related to human-machine interaction, non-invasive voice-diagnosis [5], monitoring of vocal effort, visual feedback in singing, voice-disguise, and even to improve the naturalness in speech synthesis [6].

In this paper, we present preliminary results of a new frequency-domain GPE (FD-GPE) algorithm whose conceptual approach was introduced in [7, 8]. Since according to the source-filter model, the combination in the frequency domain of the source, filter and lips/nostrils radiation is multiplicative in the magnitude domain, and additive in the phase or group-delay domains, decoupling the two aspects leads to added flexibility and accuracy, which is central to FD-GPE.

The paper is structured as follows. In section 2 we introduce representative algorithms of glottal pulse estimation that we take as reference algorithms to assess results. In section 3 we highlight the main features of the FD-GPE algorithm, explain how it was developed, and describe its operation. In section 4 we present and discuss a few results illustrating the performance of the tested algorithms, and in section 5 we summarize the main results of this paper and address next research steps.

2. REFERENCE ALGORITHMS OF GLOTTAL PULSE ESTIMATION

Several methods for estimating the glottal pulse have been proposed over the last few decades [6, 8]. Most of them fall in a category known as inverse filtering [8, 6, 9]. This approach takes advantage of the simplifying assumption that the glottal flow and the transfer function of the vocal tract are independent and thus linearly separable, which for healthy voices and modal register can be considered as quite reasonable and fairly acceptable [1]. Thus, given a short-duration voiced speech segment where local stationarity can be assumed, the power spectrum of the speech is modelled using an all-pole model which captures mainly the vocal-tract resonances, also referred to as formants. The all-pole model consists in a filter whose inverse is then used to cancel the spectral effects of the formants. Furthermore, the lip/nostrils radiation which reflects the high-pass filtering due to signal differentiation and that takes place as a consequence of the volume velocity airflow conversion to a pressure signal, is also cancelled by using an approximate signal integrator [10]. These two inverse filtering operations implement a deconvolution process leading to an estimate of the glottal source signal including the glottal pulses. A frequently cited algorithm that implements such deconvolution in an iterative and adaptive manner is IAIF [3, 4]. IAIF is implemented in an open-source Matlab toolkit named APARAT [11]. We have used APARAT as a reference algorithm.

It is known however that inverse filtering results are strongly affected by problems due to all-pole modelling and approximate signal integration. In fact, the all-pole model is naturally adapted to modelling resonances (or pole effects) and therefore performs poorly with anti-resonances (or zero effects), which are very important to accurately model nasalized vowels as well as fricative sounds. Another disadvantage of all-pole modelling that is especially problematic for high pitched harmonic sounds, is the 'harmonic locking ef-

This work was supported by the Portuguese Foundation for Science and Technology, an agency of the Portuguese Ministry for Education and Science, under research project PTDC/SAU-BEB/14995/2008. URL: http://www.fe.up.pt/∼voicestudies, last accessed on June 20th 2014.

fect' that arises because the pole locations tend to be 'locked' to the frequency of the harmonics of a voiced speech signal [6]. In addition, phase effects are automatically handled (but also constrained) by the all-pole model and, therefore, are not effectively controlled [8] which represents a lack in flexibility.

Another important glottal pulse estimation algorithm is based on spectral decomposition of the Z-transform of a voiced signal, in an anticausal part and a causal part [6]. The former corresponds to the open-phase of the glottal source, and the latter corresponds to the combination of the glottis closure and vocal tract filter. The identification of the glottal closure instants (GCI) is thus required to separate the anticausal and causal signals. The algorithm takes the Z transform of a segment of GCI-synchronized voiced speech and the roots (zeros) are computed. Three sets of zeros are identified. The set of zeros having modulus equal to one represents the impulse train underlying the speech periodicity (i.e. F0, or pitch). The set of zeros having modulus larger than one, is due to the anticausal part of the voice source. The set of zeros having modulus less than one, is due to the causal part of the voice source. Therefore, using DFT and IDFT transformation for each one of these groups, allows to estimate the glottal source and the vocal tract filter [6]. The algorithm is computationally intensive but a recent improvement named as Complex Cepstrum-based Decomposition (CC), despite requiring efficient phase unwrapping, has significantly reduced the complexity [6]. A Matlab implementation of these ZZT (Zeros of the Z Transform) and CC algorithms (ZZT-CC), as suggested to us by the author, was also used as another reference glottal pulse estimation approach. An additional Matlab function, identified as `dypsa()` and available in the popular Voicebox Matlab toolbox, was used to obtain the CGI values.

3. A NEW FREQUENCY-DOMAIN ALGORITHM TO GLOTTAL PULSE ESTIMATION

Our frequency domain approach to GPE is based on the frequency-domain analysis-synthesis framework that was described in [7]. This processing framework includes three main features that are central to the FD-GPE algorithm [7]:

1. it is able to perform accurate signal integration or signal differentiation in the frequency domain,
2. it is able to selectively analyse and resynthesize the sinusoidal/harmonic content of the signal, with or without modification,
3. it permits independent magnitude and phase signal manipulation, which adds flexibility in the processing of the magnitude and phase contributions of both glottal source and vocal tract filter (VTF).

Ultimately, these features motivated us, as anticipated in [7], to investigate the magnitude and phase characteristics of real voiced vowels, in an attempt to find more realistic models than the idealized Liljencrants-Fant (LF) or Rosenberg models for example [12]. This research provided valuable insight concerning not only the magnitude, but also the phase characterization of the glottal source periodic component of the speech. In the following we address this result and then we address the FD-GPE algorithm.

3.1. On the characterization of the glottal source

We have concluded in [12] that when the harmonics of the glottal source are analysed using Normalized Relative Delays (NRDs), which consist of a phase-related feature [7, 13], then the unwrapped NRDs of the harmonics pertaining to the periodic part of the signal captured near the vocal folds, are speaker specific and can be estimated from the NRDs of the voice harmonics captured outside the mouth, using a compensation function. Concerning magnitude characterization, physiological data suggested a realistic 'average' model may be obtained in the form of a hybrid LF-Rosenberg model [12].

Thus, in synthesising only the harmonic part of the glottal source, which leads to the desired glottal pulse signal, it is only required to cancel the magnitude effect of the VTF, since the phase component is handled by the estimated NRDs as described.

3.2. The FD-GPE algorithm

A simplified block diagram of the FD-GPE algorithm is represented in Fig. 1. A region of a voiced sound is first analysed so as to extract harmonic information and a smooth spectral envelope model (ceps). Harmonic information includes the frequencies, magnitudes (mS) and phases of all sinusoidal components that are harmonics of a fundamental frequency. Using the frequencies and phases, the NRDs are extracted (nrdS) using the approach described in [13, 8]. Based on the results presented in [12], a compensation function (nrdF) is available that allows to estimate the NRDs of the signal near the vocal folds (i.e. the glottal excitation, nrdG). This compensation function is a linear function in the unwrapped NRD domain and suggests that the acoustic signal delay between vocal folds and the region outside the mouth, predominates over non-linear phase effects due to the vocal tract filter. This combination of the magnitude and NRD characteristics due to the glottal source and vocal tract filter, is illustrated in Fig. 2. A smooth spectral envelope model (mF) that is obtained using a 16-coefficient real cepstrum, is used to model both the vocal tract resonances and anti-resonances. The dB difference between this model and the exact magnitudes of all sinusoids, is then added to the magnitude of the hybrid (prototype) glottal

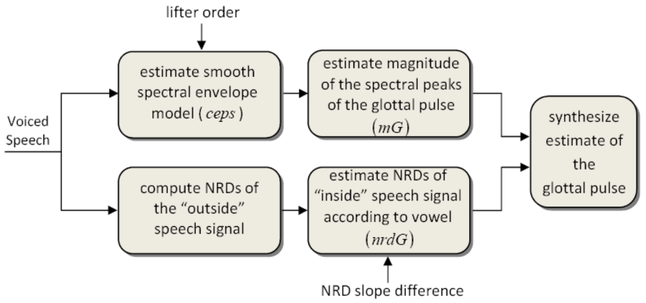
lifter order
Voiced Speech
estimate smooth spectral envelope model (*ceps*)
estimate magnitude of the spectral peaks of the glottal pulse (mG)
synthesize estimate of the glottal pulse
compute NRDs of the "outside" speech signal
estimate NRDs of "inside" speech signal according to vowel ($nrdG$)
NRD slope difference

Fig. 1. *Simplified block diagram of the GPE algorithm.*

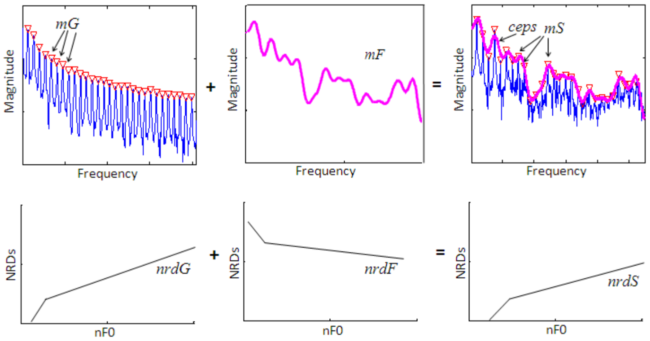

Fig. 2. *Illustration of the combination of the source and filter characteristics explaining a voiced sound. nF0 denotes harmonics of F0.*

pulse model so as to obtain the magnitudes of the estimated glottal source harmonics (mG).

Thus, as illustrated in Fig. 1, the modified magnitudes (mG) and the compensated NRDs (nrdG) of all the harmonics, are then combined to synthesize the periodic part of the glottal excitation (i.e. the glottal pulses). This operation relies on an accurate sinusoidal/harmonic synthesis that is built-in in the frequency-domain processing framework and that also implements signal integration, as described in [7, 12].

4. RESULTS AND DISCUSSION

In order to assess the quality of the FD-GPE results, we have used four synthetic signals and eight real signals. Two reference algorithms, as discussed in section 2, are included for comparison: IAIF and ZZT-CC. The former is able to deliver an estimate of the glottal pulse signal whereas the latter only provides a prototype estimate of the glottal pulse derivative. Therefore, explicit signal comparisons and quantitative evaluations are possible in the first case, while in the second case only a qualitative evaluation is made.

4.1. Tests using synthetic signals

In order to generate the synthetic signals, we have used the Voicebox Matlab toolbox and the LF model for the glottal source signal. Two Portuguese vowels, /a/ (as in ‘bath’) and /i/ (as in ‘heed’), have been synthesized using a 6th-order all-pole model and using two fundamental frequencies simulating male and female voices, 110 Hz and 300 Hz, respectively. As an example, Fig. 3 represents for the synthesized male /a/ vowel, the reference LF model, its derivative, as well as the estimated glottal signals using IAIF and FD-GPE. It can be seen that these two signals are rather similar, and the same conclusion is valid for the remaining three synthetic vowels. The results suggest that the glottal pulse estimates by both algorithms do not exhibit a clear closed phase of the glottal pulses, which needs further research and clarification. When comparing the estimates by IAIF and ZZT-CC of the glottal pulse derivative, it is clear that the latter have less ripple and seem to perform better when F0 is higher.

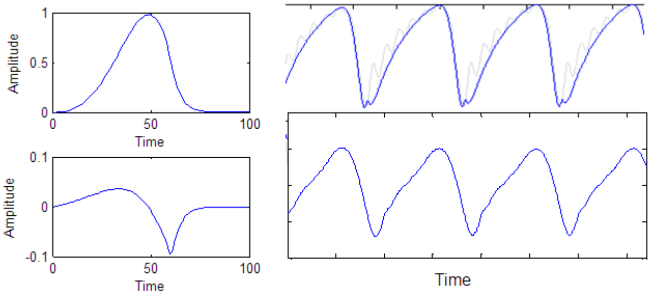

Fig. 3. *Upper left: ideal LF glottal model. Lower left: derivative of the LF model. Upper right: IAIF estimated glottal signal. Lower right: FD-GPE estimated glottal signal.*

Table 1 presents the SNR evaluating the ratio (in deciBel) between the power of the ideal reference glottal signal, and the power of the difference between this reference and the estimated signals according to the IAIF or FD-GPE algorithms, after proper signal alignment and scaling. Table 1 confirms that results between the two algorithms are quite comparable, and that performance is better when F0 is higher, which does not confirm the idea found in the literature that IAIF is less accurate in the case of female speech.

Table 1. *SNR of the estimated glottal signals according to the IAIF and FD-GPE algorithms.*

SNR (dB)	**Male /a/**	**Male /i/**	**Female /a/**	**Female /i/**
IAIF	6,38	6,09	9,80	8,72
FD-GPE	7,30	8,72	9,75	8,87

4.2. Tests using natural signals

In [8] we present two sets of results involving natural speech signals. The first set includes four voiced vowels and no ‘reference’ glottal signal is available. Therefore, only a qualitative comparison is possible among the glottal pulse estimations (or their derivatives) provided by the IAIF, ZZT-CC, and FD-GPE algorithms. Is has been observed that the IAIF and FD-GPE provide quite similar glottal pulse estimations, which is in line with conclusions discussed in section 4.1. On the other hand, it has been concluded that the closing and closed phases of the estimated glottal pulses, differ significantly from the shape associated with idealized models such as the LF model. This can either be explained by the fact that the estimation algorithms are not accurate, which however is not entirely plausible since the algorithms are strongly different in their processing approach, or perhaps the true form of the glottal pulses are not realistically represented by the idealized glottal pulse models.

Concerning the second set of results, five voiced vowels were used that are included in the database that we describe in [12] and that motivated the proposal of a hybrid LF-

Rosenberg model. One interesting advantage of this database is that two versions of each vowel are available and that correspond to synchronized records (using special microphones) of the vowel signal captured near the glottis, as well as the vowel signal captured outside the mouth. Thus, a reference signal is available that we take as a relevant acoustic evidence of the glottal pulse. An example is illustrated in Fig. 4 that corresponds to vowel uttered by a male speaker. This figure also represents the estimated glottal pulses according to the IAIF and FD-GPE algorithms. We were expecting that the reference signal would be similar to the glottal pulse derivative since the pressure signal captured by he microphone is the derivative of the volume velocity airflow [2]. Instead, for all records, we found a reference signal that better resembles the shape of the glottal pulse than the shape of its derivative. This is a topic for further research. In any case, and as Fig. 4 suggests, the glottal pulse estimated using FD-GPE is closer to the reference signal than the glottal pulse estimated using IAIF is. In fact, on average, the SNR in the first case is 3.6 dB better than in the latter case [8]. Observations of the reference signal as well as of the glottal pulse derivative signals using ZZT-CC, also reinforced our remarks and discussion above concerning the closing and closed phases of the glottal pulse.

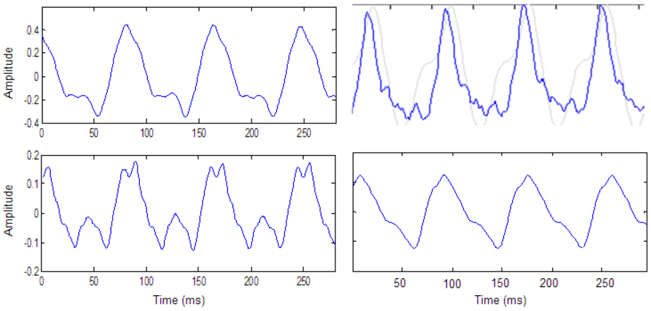

Fig. 4. *Upper left: signal captured near the glottis. Lower left: signal captured outside the mouth. Upper right: IAIF estimated glottal signal. Lower right: FD-GPE estimated glottal signal.*

5. SUMMARY

We described a frequency-domain glottal pulse estimation algorithm that takes advantage of the sinusoidal and noise diversity of a voiced sound, and that decouples de magnitude and phase contributions of both source and filter parts of voice production. Results were presented using synthetic signals as well as natural voice signal, and taking as a reference two glottal pulse estimation algorithms, IAIF and ZZT-CC. Despite the difficulty of assessing performance in the case of natural voice sounds since the 'ground-truth' is not available, results have highlighted that the estimates provided by IAIF and FD-GPE are quite comparable, which is interesting although surprising since the underlying algorithms are strongly different. Results have also suggested that all tested algorithms have an inherent difficulty dealing with the closing and closed phases of the glottal pulses, which differ significantly from established idealized glottal pulse models. Further research efforts will be directed to address this difficulty, to achieve real-time glottal pulse estimation, and to artificially implant voicing on whisper.

6. REFERENCES

[1] G. Fant, *Acoustic Theory of Speech Production*. The Hague, 1970.

[2] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 7, no. 5, pp. 569–586, September 1999.

[3] P. Alku, "An automatic method to estimate the time-based parameters of the glottal pulseform," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1992, pp. II–29–II–32.

[4] L. Lehto, M. Airas, E. Bjokner, J. Sundberg, and P. Alku, "Comparison of two inverse filtering methods in parametrization of the glottal closing characteristics in different phonation types," *Journal of Voice*, vol. 21, no. 2, pp. 138–150, 2007.

[5] K. Murphy, "Digital signal processing techniques for application in the analysis of pathological voice and normophonic singing voice," Ph.D. dissertation, Facultad de Informática (UPM), Spain, 2008.

[6] T. Drugman, "Advances in glottal analysis and its applications," Ph.D. dissertation, University of Mons, Belgium, 2011.

[7] S. Dias, R. Sousa, and A. Ferreira, "Glottal inverse filtering: a new road-map and first results," in *Speech Processing Conference*, June 2011, tel-Aviv, Israel. [Online]. Available: http://www.fe.up.pt/∼voicestudies, last accessed on June 20th 2014.

[8] S. Dias, "Estimation of the glottal pulse from speech or the singing voice," 2012, MSc dissertation.

[9] J. Walker and P. Murphy, "A review of glottal waveform analysis," *Lecture Notes in Computer Science - Progress in Nonlinear Speech Processing*, vol. 4391, pp. 1–21, 2007, springer-Verlag.

[10] H. R. Javkin, N. A. Barroso, and I. Maddieson, "Digital inverse filtering for linguistic research," *Journal of Speech and Hearing Research*, vol. 30, pp. 122–129, 1987.

[11] M. Airas, "TKK aparat: Enviroment for voice inverse filtering and parameterization," *Logopedics Phoniatrics*, vol. 33, no. 1, pp. 49–64, 2008.

[12] S. Dias and A. Ferreira, "A hybrid LF-Rosenberg frequency-domain model of the glottal source," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2013.

[13] R. Sousa and A. Ferreira, "Importance of the relative delay of glottal source harmonics," in *39th AES International Conference on Audio Forensics - practices and challenges*, 2010, pp. 59–69.