

A HYBRID LF-ROSENBERG FREQUENCY-DOMAIN MODEL OF THE GLOTTAL PULSE

Sandra Dias, Aníbal Ferreira

Department of Electrical and Computer Engineering
University of Porto - Faculty of Engineering, Porto, Portugal
ajf@fe.up.pt

ABSTRACT

In this paper we describe innovative advances to the design of a new frequency-domain algorithm to glottal source estimation whose conceptual approach we have reported recently [1]. Those advances result from accurate sinusoidal/harmonic analysis and synthesis of two concomitant acoustic signals: the glottal source signal captured near the vocal folds, and the corresponding voiced signal captured outside the mouth. We describe the experimental procedure which was performed by an ORL specialist using a rigid video-laryngoscope and two tiny and high-quality microphones. Six subjects have participated in the tests and records were made for vowels /a/ and /i/. The data analysis allowed us to conclude on the magnitude and on the phase-related NRD features of the glottal source signal. In addition, a new frequency-domain glottal pulse model combining features of the Liljencrants-Fant and Rosenberg models has been devised that is a better match to the observed data. The derivatives of the three models are obtained using accurate frequency-domain processing. The paper concludes with next research steps.

Index Terms— Glottal source modelling, Normalized Relative Delays.

1. INTRODUCTION

The non-invasive estimation of the glottal source signal is gaining importance in practical application scenarios including diagnosis of the vocal folds [2], speaker recognition [3], emotion analysis [4, 5], and natural speech synthesis [6].

Most approaches estimating the glottal source are time-domain based and involve inverse filtering of the speech signal with the purpose to cancel the effect of the vocal tract filter (VTF), and to cancel the lips/nostrils radiation effect. Time-domain approaches rely on specific assumptions and approximations which constrain the validity or the quality of the glottal source estimation. For example, it is typically assumed that the VTF can be modelled as an all-pole filter. This assumption is more motivated by the analytical convenience of LPC modelling [7] than by a match to observed data describing the VTF alone. In particular, the all-pole model constrains the magnitude and group delay response of the VTF, making difficult the modelling of nasal sounds for example. On the other hand, the lips/nostrils radiation is typically modelled as a first-order discrete-time differentiation [8] and reversing its effect in the time domain is

problematic and inaccurate. In [1] we have shown that it is possible to overcome the associated inaccuracies by reversing the radiation effect in the frequency domain.

Several time-domain glottal source estimation algorithms consistently reveal critical problems with high pitched and pathological voices [9]. On the other hand, frequency domain approaches to glottal source estimation use the zeros of the Z-Transform (ZZT) [10] to perform a spectral decomposition of the voice signal into an anti-causal component denoting the open phase of the glottal cycle, and a causal component denoting the closed phase of the glottal cycle. A critical aspect of this approach, other than the considerable computational complexity, is its dependency of an accurate time-domain localization of the glottal closure instants (CGI). Recent improvements of this approach have relaxed the complexity issue but still rely on accurate time localization of CGI as well as on phase unwrapping [6].

In [1] we have presented the concept of a completely new frequency-domain approach to glottal source estimation which does not depend on GCI localization and that relies instead on accurate sinusoidal/harmonic analysis and synthesis, as well as on new phase-related features consisting of the Normalized Relative Delays (NRDs) of the harmonics pertaining to the glottal source [11]. NRDs describe the normalized temporal shift of each harmonic with respect to the onset of the fundamental. In this paper we describe innovative advances to the implementation of that approach that result from a set of experiments conducted by an Otorhinolaryngologist (ORL). A rigid video laryngoscope was used as in a conventional laryngoscopy exam, in order to record two time-aligned acoustic signals: a glottal source signal captured as close as possible to the vocal folds, and the corresponding voiced signal captured outside the mouth. The main purpose in obtaining these signals was to model the spectral magnitude and NRDs of the glottal source signal (the main focus of this paper), as well as to find a model for the frequency response of the VTF corresponding to each one of the vowels /a/ and /i/. These models are a pre-requisite to implement the approach devised in [1].

The remaining of this paper is structured as follows. In section 2 we describe the physiological signal acquisition procedure. In section 3 we characterize the source signals obtained near the glottis, in terms of spectral magnitude and NRDs. As a result of this characterization and analysis, we present in section 4 a new hybrid Liljencrants-Fant/Rosenberg frequency model of the glottal pulse which is a better match to the experimental data and that has interesting properties, namely the fact that it is independent of the fundamental frequency (usually referred to as F0 or pitch). Finally, section 5 summarizes the main results of this paper and addresses next development steps.

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://gnomo.fe.up.pt/~voicestudies/artts/>, last accessed on May 17th 2013.

2. PHYSIOLOGICAL SIGNAL ACQUISITION

Assuming the source-filter model of voice production [12], in order to understand a voiced signal (*e.g.* corresponding to a vowel utterance) captured through a microphone outside the mouth, it is necessary to know the nature of the source signal (*i.e.*, the glottal excitation signal), as well as the VTF, including the lips/nostrils radiation effect. However, the glottal source signal is not directly observable. Several ideal (*i.e.*, mathematical) models exist [13], such as the Rosenberg and LF models [12, 13], but their correspondence to physiological data is not clear. We therefore devised a technique, after consulting several specialists concerning viability, feasibility, ethical and technical aspects, in order to capture the source signal as close as possible to the vocal folds.

As a consequence, an experimental procedure was carefully planned and conducted by an ORL professional, whose aim was to capture the two time-aligned acoustic signals as mentioned above. For this purpose we attached a tiny and high-quality microphone (of the ear-worn type) at the tip of a rigid video-laryngoscope. An identical ear-worn microphone was used to capture the voiced signal outside the mouth. The recording sessions took place during conventional video-laryngoscopy examinations. The ORL professional acted so as to capture the glottal source signal as close as possible to the vocal folds, while insuring safety and ethical conditions for all volunteer subjects. Fig. 1 illustrates the experimental procedure as well as the equipment involved. The equipment used in this ex-

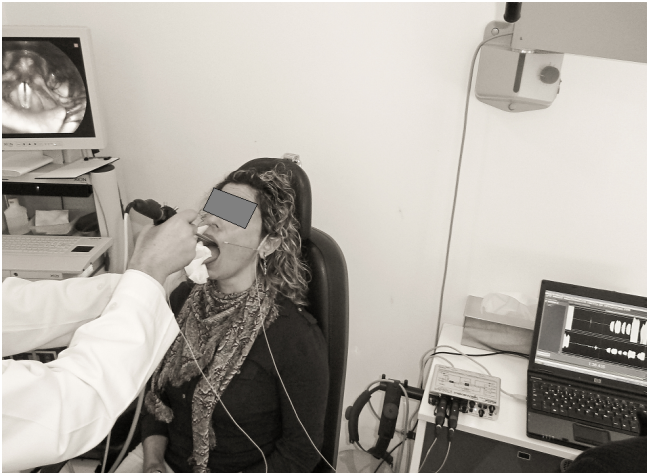


Figure 1: *Laryngoscopy examination and voice signal acquisition. One microphone is visible outside the mouth and a similar microphone is attached at the tip of the rigid laryngoscope (not visible) near the vocal folds (represented in the screen at the top-right side of the image). The A/D interface and audio recorder are displayed at the bottom-right side of the image.*

perimental procedure was i) a rigid video-laryngoscope with 7mm diameter and length of 180mm (Xion); ii) two ear-worn omnidirectional pre-polarized condenser microphones, with a frequency response from 20Hz to 20kHz, maximum of 143dB SPL, and a 3.3mm diameter microphone capsule (Sennheiser Ear Set 1 microphones); iii) two phantom power adaptors (XLR to 3.5mm mini-jack plug) for pre-polarized condenser microphones (Sennheiser MZA-900P); iv) a stereo 24-bit/96kHz A/D and D/A interface with phantom power, XLR inputs and USB PC connector (Cakewalk-Roland UA-25EX); and v) Adobe Audition audio recorder/editor (Adobe).

Six adult volunteer subjects (Table 1), with no voice disorders, participated in the examination and record sessions, and were asked to utter vowels /a/ and /i/. Despite the fact that the same procedure

Table 1: *Characterization of the volunteer subjects.*

Subject	Gender	Age	Occupation
AF	Male	48	Teacher
RT	Male	54	Teacher/Tenor
PP	Male	22	Speech Therapist
SD	Female	31	Teacher
SF	Female	33	Teacher/Speech Therapist
CK	Female	49	Singer

and equipment was used for all subjects, differences in the signal acquisition near the vocal folds were noticed, and were concluded to be subject dependent. The main reasons are anatomical and subject collaboration. In order to minimize these aspects, several records were made for the same subject so as to select the ‘best’ signals. It was also concluded that the quality of the acoustic signal captured near the vocal folds depended significantly on the proximity of the microphone to the epiglottis. If too far, the signal reveals a noticeable influence of the vocal tract resonances. After the recordings and using the audio editor, we selected for each subject and vowel, two representative segments, each about 2 seconds long, of the time-aligned acoustic signals. These signals represent the data in our database. It consists of $(6 \times 2 \times 2) = 24$ stereo, 16-bit, WAV files. The original recording sampling frequency was 48 kHz but the files were down-converted to 22050 Hz in order to facilitate subsequent processing, namely in terms of spectral analysis.

3. DATA ANALYSIS FOR GLOTTAL SOURCE MODELLING

For each one of the 24 acoustic signals obtained near the glottis, we characterized in detail the magnitude and NRDs of the harmonics. Concerning magnitude, we found first the magnitude of individual harmonics for each frame of the source signal, as illustrated in Fig. 2. Then, for each vowel record and speaker, we determined

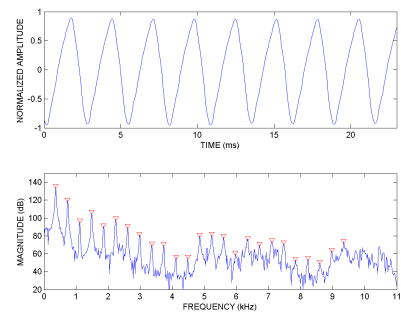


Figure 2: *Waveform of the signal captured near the glottis for vowel /i/ by a male subject (upper figure) and the corresponding magnitude spectrum, harmonics are marked (lower figure).*

the average normalized magnitude (relative to the magnitude of F0) of all detected harmonics. From this information we obtained the corresponding Least Squares Regression model in the logarithmic scale with the purpose to compare the resulting slope with the -12 dB/octave reference. This slope reference is usually considered in the literature as the natural decay of the glottal pulse spectrum [12].

Interestingly, we have found this is approximately the slope of the Rosenberg glottal pulse model, as we detail in the next section.

Table 2 presents the spectral decay results according to gender and vowel. These results have been obtained using on average 16 harmonics for vowel /a/ and 11 harmonics for vowel /i/. This difference is related to the known fact that vowel /i/ has an intrinsic pitch which is higher than that of vowel /a/. It can be concluded that results per vowel are quite consistent between genders, and that the standard deviation reveals there are no strong differences among speakers of the same gender. Considering both genders, the average slope for vowel /a/ is about -10 dB/octave and, for vowel /i/, is about -14 dB/octave. Thus, according to our data, the average slope for vowel /a/ is lower than the -12 dB/octave reference, while the average slope for vowel /i/ falls between the slope¹ of the Rosenberg model (about -12 dB/octave) and that of the LF model (about -16 dB/octave).

Table 2: Average and STDev of the spectral decay of the source harmonics, in dB/octave, as a function of gender and vowel.

	AVG /a/	STDev /a/	AVG /i/	STDev /i/
males	-9,64	2,33	-13,73	2,87
females	-9,93	3,35	-13,67	2,26

A final note regards NRD models. The recorded signals have negative polarity, which has an impact in the behaviour of the NRDs, as already detected in [11]. Analysing the results of the NRD parameters for different vowels and subjects, as illustrated in Fig. 3 for one particular subject, vowel and data frame, it is clear that in most cases the unwrapped NRDs can be approximated by a line, however its slope is strongly subject dependent. We have

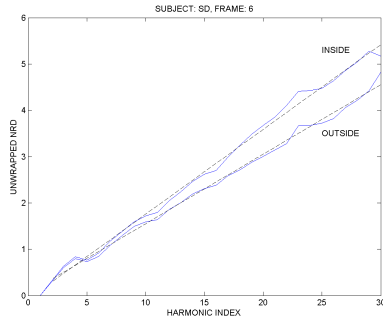


Figure 3: Unwrapped NRDs of two time-aligned signal frames captured 'inside' and 'outside' the mouth when a female subject was uttering vowel /a/. The dashed lines represent the corresponding linear approximation models.

observed slopes between almost 0/30 and about 9/30, with most frequent cases falling around 3/30 and 6/30. We therefore conclude that NRDs express idiosyncratic information and that average models can not be obtained per gender or vowel.

4. A HYBRID LF/ROSENBERG MODEL

We have concluded in the previous section that using physiological data, models for the spectral decay of source harmonics could be found and, despite the fact that unwrapped NRDs can be approximated by a line, its slope is highly subject dependent and therefore

¹The Rosenberg and LF models have been synthesized using the `glotros` and `glotlf` Matlab functions (with default parameters) included in the Voicebox toolbox.

representative models can not be established. However, as indicated in section 2, our approach to glottal source estimation [1] requires a default model of the glottal pulse. We have thus analysed two popular models. Using the (publicly available) Voicebox Matlab toolbox, we have synthesized ideal LF glottal pulses, as well as Rosenberg glottal pulses. A careful analysis has revealed that the spectral slope of the Rosenberg model is -11,84 dB/octave, while that of the LF model is -16,43 dB/octave. The Rosenberg model is not an option because it is not a smooth model (i.e. its derivative has a discontinuity). On the other hand, the spectral decay of the LF model is not a good match to our physiological data. Using the normalized magnitudes of the first 30 harmonics of each model, as well as the corresponding unwrapped NRD coefficients, we decided to build a hybrid model. Fig. 4 depicts the normalized magnitudes (relative to the magnitude of F0) of the first 30 harmonics pertaining to the LF and to the Rosenberg glottal pulse models. The normalized magnitudes of the hybrid model are just the average of the two models, as also represented in Fig. 4. Fig. 5 represents the un-

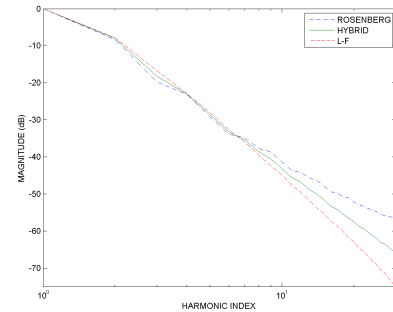


Figure 4: Normalized magnitudes of the first 30 harmonics pertaining to the LF glottal pulse model (dashed line), the Rosenberg model (dash-dotted line) and a hybrid model (solid line).

wrapped NRDs of the LF model as well as its linear approximation ($y = 0.0511 + 0.0849x$, where x represents the index of the harmonic, $x \geq 2$ since by definition the NRD of F0 is always zero), and the unwrapped NRDs of the Rosenberg model as well as its linear approximation ($y = 0.2412 + 0.0090x$). We have used dif-

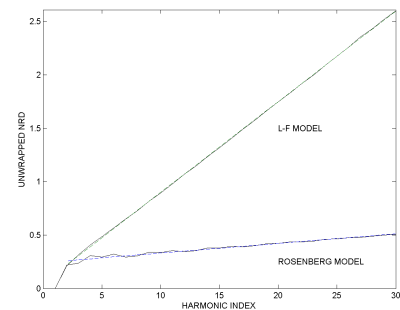


Figure 5: Unwrapped NRDs of the LF and Rosenberg glottal pulse models (solid lines) as well as the corresponding linear regression models (dashed lines).

ferent combinations of both linear regressions in order to build an NRD model for the hybrid glottal pulse model. The smoothest result was obtained for the LF NRD linear model. The fact that it fits quite accurately the unwrapped NRD coefficients of the LF model, as can be seen in Fig. 5, helps to explain this outcome. Fig. 6 illustrates a time-domain representation of all three models, and Fig.

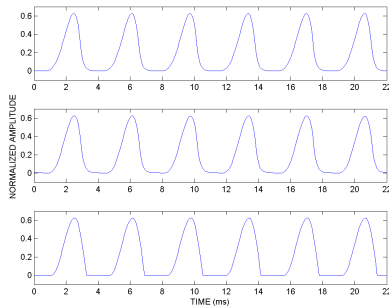


Figure 6: Time-domain glottal pulse models: LF model (top), hybrid model (center), and Rosenberg model (bottom).

7 illustrates their derivatives. It should be noted that the derivative

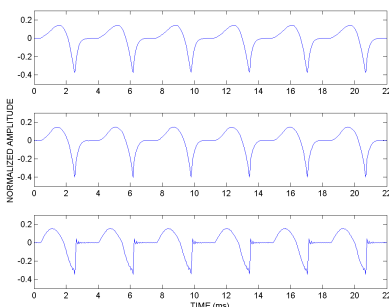


Figure 7: Derivative of glottal pulse models: LF model (top), hybrid model (center), and Rosenberg model (bottom).

signals have been obtained by acting exclusively in the frequency domain, using our analysis-synthesis processing framework which is detailed in [1]. In fact, since derivation in the time domain corresponds to a multiplication by $j\Omega$ in the frequency domain, we just shift the phases of all harmonic sinusoids by $\pi/2$ before synthesis, and we scale their magnitude by $|\Omega|$ at the exact frequency of each harmonic sinusoid since it is not constrained to the discrete frequencies of the analysis-synthesis filterbank [14]. It is interesting to note in Fig. 7 a small ripple around the discontinuity in the derivative of the Rosenberg model (as expected due to the Gibbs phenomenon) and it is interesting to conclude, using both Figs. 6 and 7, that the hybrid model approaches the LF model during the closed phase of the glottal pulse, and approaches the Rosenberg model during the open phase of the glottal pulse. A very convenient property of normalized spectral magnitudes and NRDs of harmonics, is that they allow to completely define a time waveform shape, independently of its F0, time-shift or amplitude.

5. CONCLUSION

We have addressed recent advances to the design of a new frequency-domain approach for glottal source estimation, notably the specification of a new glottal pulse prototype. These results were made possible because of special voice data acquisition conditions which were described. Next research steps will be devoted to finalise the development of an analysis-by-synthesis algorithm allowing the joint estimation of glottal source and vocal tract filter, as in [2].

6. ACKNOWLEDGMENTS

The authors would like to thank all six volunteer subjects and Dr. Seabra da Rocha for the kind collaboration and skilled technical assistance on laryngoscopy examination and data acquisition.

7. REFERENCES

- [1] 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://gnomo.fe.up.pt/~voicestudies/artts/publications.html>, last accessed on May 17th 2013.
- [2] 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. [Online]. Available: <http://oa.upm.es/1079/>, last accessed on May 17th 2013.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] T. Drugman, “Advances in glottal analysis and its applications,” Ph.D. dissertation, University of Mons, Belgium, 2011.
- [7] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice-Hall, 1978.
- [8] H. R. Javkin, N. A. Barroso, and I. Maddieson, “Digital inverse filtering of linguistic research,” *Journal of Speech and Hearing Research*, vol. 30, pp. 122–129, 1987.
- [9] M. Airas, “TKK aparat: Environment for voice inverse filtering and parameterization,” *Logopedics Phoniatrics*, vol. 33, no. 1, pp. 49–64, 2008.
- [10] B. Bozkurt, B. Duval, C. d’Alessandro, and T. Dutoit, “Zeros of z-transform representation with application to source-filter separation in speech,” *IEEE Signal Processing Letters*, vol. 12, no. 4, pp. 344–347, 2005.
- [11] 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.
- [12] G. Fant, *Acoustic Theory of Speech Production*. The Hague, 1970.
- [13] —, “Glottal flow: models and interaction,” *Journal of Phonetics*, vol. 14, no. 3/4, pp. 393–399, 1986.
- [14] A. J. S. Ferreira, “Combined spectral envelope normalization and subtraction of sinusoidal components in the ODFT and MDCT frequency domains,” in *2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 21–24 2001, pp. 51–54.