

Methods for Estimation of Glottal Pulses Waveforms Exciting Voiced Speech

Milan Boštík, Milan Sigmund

Department of Radio Electronics
Brno University of Technology, Czech Republic
bostik|sigmund@feec.vutbr.cz

Abstract

Nowadays, the most popular techniques of the speech processing are the recognition of all kinds (the speech, the speaker and the state of speaker recog.) and the text-to-speech synthesis. In both these domains, there are possibilities to use the glottal pulses waveforms. In the recognition techniques we can use them for the vocal cords description and then use it for the classification of speaker's state (physiological or mental state) or for the classification of a speaker. In the text-to-speech techniques we can use them for the speech timbre changing. This paper describes some methods for obtaining of glottal pulses waveforms from recorded speech. There are several results obtained by application of described methods.

1. Introduction

Human speech arises when the air flows through glottal gap (it is situated near the entrance of larynx), buccal cavity, oral cavity and nasal cavity. Lips and nose radiate the speech out of body to a free space. Vocal cords are elastic muscles and human can stretch them more or less and by this manner can determine speed of their oscillation. Frequency of vocal cords oscillation determines a level of fundamental vocal cords frequency of speech that is one of the basic parameters of human speech. In Fig. 1 there is shown arrangement of vocal cords at phonation [1].

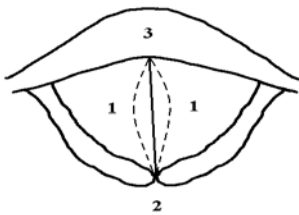


Figure 1: The basic arrangement of vocal cords at phonation: 1 – vocal cords, 2 – transverse muscle, 3 – part of epiglottis.

The dashed lines describe open phase, the solid line describes closed phase. If the pressure of air (goes from lungs) starts to impress to close vocal cords (solid line), vocal cords become open (dashed line). After strength escape of air the vocal cords became closed and the situation is repeated. Through the glottal gap a various quantity of air in various times flows and the quantity corresponds to size of glottal gap. The main goal of this contribution is to introduce the glottal pulses estimation and to show their waveforms, which arises by vocal cords opening [2], [3].

2. Glottal pulses waveforms obtained by linear prediction error estimate

This method is based on simplified digital model of vocal apparatus, which is described in detail e.g. in [4], [5]. Common model for voiced speech is shown in Fig. 2.

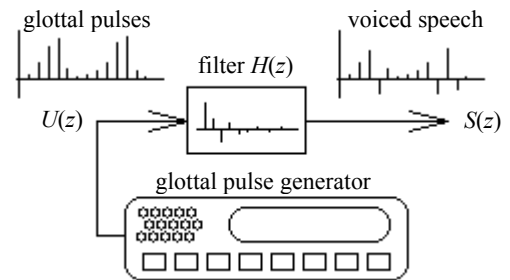


Figure 2: Speech production model for voiced speech.

In the time domain the system can be described as

$$s(k) = -\sum_{i=1}^M a_i s(k-i) + u(k) \quad (1)$$

where $s(k)$ is speech signal and $S(z)$ is its representation in z -domain (see Fig. 2), a_i are linear predictive coefficients (LPC) of predictor with order M , $u(k)$ is glottal pulse waveform and $U(z)$ is its representation in z -domain (see Fig. 2). Then the transfer of linear filter $H(z)$ can be written as

$$H(z) = \frac{S(z)}{U(z)} = \frac{1}{1 + \sum_{i=1}^M a_i z^{-i}} \quad (2)$$

Relative simplicity of obtaining excitation characteristics implies from the Eq. (2). If we describe production of speech we move from glottal pulse generator through linear filter $H(z)$ to speech signal. So, if we want to describe excitation we must move in reverse direction - through the linear filter with inverse characteristic

$$H^{-1}(z) = 1 + \sum_{i=1}^M a_i z^{-i} \quad (3)$$

So, we obtain excitation of vocal tract by filtering of speech signal by simple linear filter $H^{-1}(z)$. We must appreciate that order of the LPC predictor cannot be too oversized or too undersized, in practice $M = 8 \div 20$. Filter $H^{-1}(z)$ must suppress formant's frequencies, which characterise vocal tract and at the same time keep information about excitation of vocal tract.

We may not forget that very simple model of vocal tract is used; hence the glottal pulse waveform at the output of the filter $H^{-1}(z)$ is not so good as expected. Therefore, it is suitable to use other low pass filter, e.g. averaging of Q samples in time domain

$$K(z) = 1 + \sum_{i=1}^Q z^{-i} \quad (4)$$

Whole process of glottal pulses waveform obtaining by LPC method is shown in Fig. 3. The LPC were computed using the autocorrelation algorithm (Levinson recursion [5]).

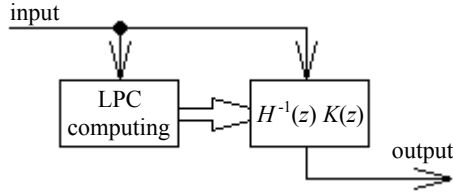


Figure 3: Block scheme for estimation of glottal pulses waveforms by linear prediction.

3. Glottal pulses waveforms obtained by cepstral coefficients

The method we describe in previous paragraph we can analyse in frequency domain too. Let us concentrate especially on filter $H^{-1}(z)$, which realises multiplication with input signal spectrum and its frequency characteristic. Because the frequency characteristic of filter $H^{-1}(z)$ is smooth, and approximates spectrum of the output signal, excitation of vocal tract is given by quickly changing part in the spectrum of input speech signal. Generally, we can obtain spectrum of glottal pulses by dividing spectrum of speech and its smoothed spectrum. In the previous paragraph we used the LPC for smoothing the spectrum. In case of this method we use cepstral coefficients for smoothing the spectrum. In comparison with LPC, spectrum peaks are more rounded. The cepstrum is defined as inverse Fourier transform of spectrum logarithm

$$c(n) = \text{IDFT}\{\log(\text{DFT}\{s(n)\})\} \quad (5)$$

Coefficients $c(n)$ with lower index n characterize formant structure of speech (slow changes in the spectrum), the coefficients with higher index n characterize glottal pulses (quick changes in the spectrum).

Whole process of glottal pulses waveform obtaining by cepstral coefficients is shown in Fig. 4.

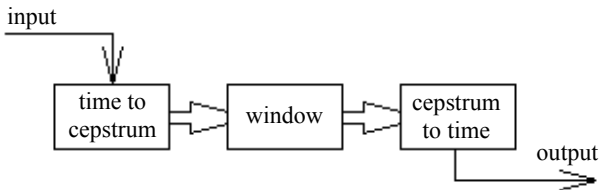


Figure 4: Block scheme for estimation of glottal pulses waveforms by cepstral coefficients.

4. Glottal pulses waveforms obtained by ARMA modeling

The ARMA modeling of vocal tract is very similar to the linear prediction modeling [6], [7]. System for the ARMA modeling can be described in time domain as

$$s(k) = -\sum_{i=1}^Q a_i s(k-i) + \sum_{j=0}^P b_j u(k-j) \quad (6)$$

where $s(k)$ is speech signal and $S(z)$ is its representation in z -domain (see Fig. 2), a_i are autoregressive (AR) coefficients of predictor with order Q , b_i are moving-average (MA) coefficients of predictor with order P , $u(k)$ is glottal pulse waveform and $U(z)$ is its representation in z -domain (see Fig. 2). Transfer function of ARMA filter $H_A(z)$ is

$$H_A(z) = \frac{S(z)}{U(z)} = \frac{\sum_{j=0}^P b_j z^{-j}}{1 + \sum_{i=1}^Q a_i z^{-i}} \quad (7)$$

and filter $H_A^{-1}(z)$ for obtaining glottal pulses waveforms is inverse to filter $H_A(z)$

$$H_A^{-1}(z) = \frac{1 + \sum_{i=1}^Q a_i z^{-i}}{\sum_{j=0}^P b_j z^{-j}} \quad (8)$$

Whole process of glottal pulses waveform obtaining by ARMA modeling is shown in Fig. 5. Coefficients a_i and b_j in Eq. (8) were computed by means of the Prony's algorithm [6].

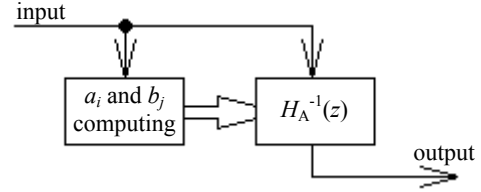


Figure 5: Block scheme for estimation of glottal pulses waveforms by ARMA modeling.

5. Pulses waveforms obtained by smoothing speech spectra

In many cases it is not necessary to obtain glottal pulses waveforms. Some time we can explore time variability of the waveform. For example in case of text-to-speech synthesis, we want to add information about speaker (timbre changing etc.) to speech. One of the methods for measuring the variability is as follows.

In Fig. 6-a there are shown two periods of speech (a vowel). Now, we are interesting about time variability between glottal pulses waveforms (between the first and the second period). So we can obtain the spectrum of the vowel using discrete Fourier transform (see Fig. 6-b).

In Fig. 7-a there is shown modified segment of the speech from Fig. 6-a. Signal in the second period substitutes the constant signal with zero's amplitude. The effect of that is that computed spectrum is smoother (see Fig. 7-b). Thus, if we divide the spectrum in Fig. 6-b and the spectrum in Fig. 7-b we obtain the spectrum of the pulse that describes time variability of the waveform (see Fig. 8).

This waveform (Fig. 8-b) we can process with using the thresholding and we can compute, for example, the ratio in

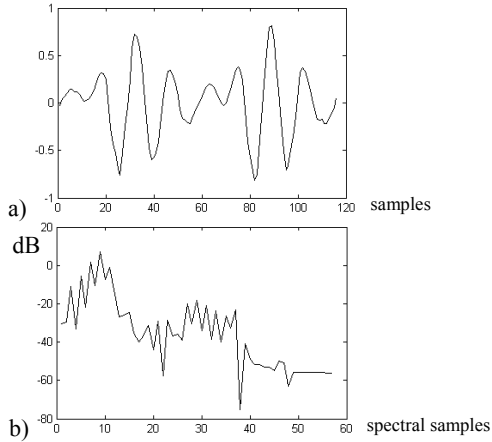


Figure 6: Segments (two periods) of speech signal - a and the spectrum - b.

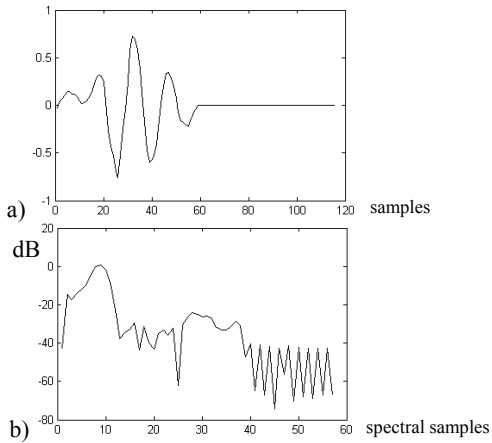


Figure 7: Modified segment of speech signal - a and the spectrum - b.

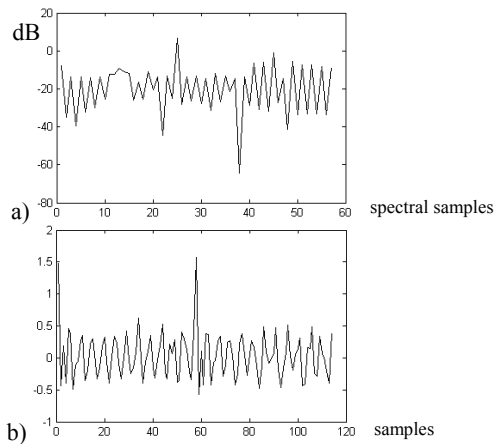


Figure 8: Spectrum of variability pulse - a and time waveform - b.

Eq. 9, where E_{above} is the energy of the signal above the threshold only and E_{below} is the energy of the signal below the threshold only.

$$R = \frac{E_{above}}{E_{below}} \quad (9)$$

It is possible to estimate also another parameters that can describe the variability pulse [8], [9].

6. Obtained results

All the described methods were applied on the speech data – only voiced segments of speech. Data were obtained from the American database SUSAS (Speech Under Simulated and Actual Stress), because the algorithms described above are suitable for recognition of mental state of speaker. In the following figures (Fig. 9) there are shown glottal pulses waveforms that were obtained from the same speech data by particular described methods.

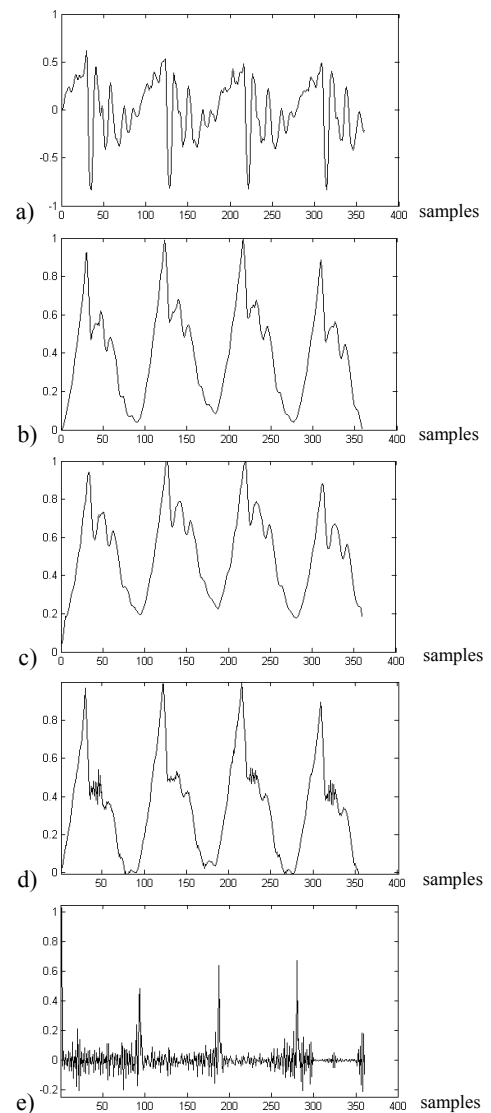


Figure 9: Speech waveform - a; glottal pulses waveforms obtained by various method: linear prediction - b; cepstral coefficients - c; Prony's algorithm - d; smoothing spectrum - e.

The results shown in Fig. 9 were obtained using 11 coefficients for linear prediction method, 11 cepstral coefficients for cepstrum, 2 coefficients for MA model and 11 coefficients for AR model in the Prony's method. It is necessary to remark that the waveforms illustrated in Fig. 9-b,c,d are the output waveforms from the particular outputs shown in Fig. 3, Fig. 4 and Fig. 5 after passing through filter

$$F(z) = \frac{1}{1-z^{-1}} \quad (10)$$

The presented approaches were also successfully applied together with methods presented in [10] for recognition of mental state of speakers. The first results can be seen in [11]. The methods are suitable especially for simulated stress such as in the SUSAS database that was used in the experiments.

7. Conclusions

Four methods for estimation of glottal pulses waveforms are presented in this contribution. We can see that these waveforms are very similar. Resulting pulses waveforms obtained by linear prediction and cepstrum analysis are smoother than waveforms obtained by Prony's algorithm. In the obtained pulses can be found some parameters based on geometrical form in the time domain. Such parameters are useful for special voice analysis, for example detection of pathological voices or emotion states of speaker.

8. References

- [1] Hála, B., *Fonetika v teorii a v praxi*, SPN, Praha 1975.
- [2] Baken, R. J., *Clinical Measurement of Speech and Voice*, Singular Publishing Group, Inc. New York, 1996.
- [3] Boyanov, B. and Baudoín, G., "Acoustical Analysis of Pathological Voice", *Proc. of the 3rd Slovenian-German Workshop Speech and Image Understanding*, Ljubljana, 1996, p. 157-166.
- [4] Psutka, J., *Komunikace s počítačem mluvenou řečí*, Academia, Praha 1995.
- [5] Sigmund, M., *Voice Recognition by Computer*, Tectum Verlag, Marburg, 2003.
- [6] <http://www.mathworks.com/access/helpdesk/help/toolbox/signal/spectop8.shtml>
- [7] El-Jardouli, A. and Makhoul, J., "Discrete All-Pole Modeling", *IEEE Transaction on Signal Processing*, vol. 39, No. 2, 1991, p. 411-418.
- [8] Alku, P. and Wilkman, E., "A Frequency Domain Method for Parameterisation of the Voice Source", <http://www.asel.udel.edu/icslp/cdrom/vol3/076/a076.pdf>
- [9] Strik, H., "Automatic Parameterisation of Voice Source Signals: a Novel Evaluation Procedure is Used to Compare Methods and Test the Effects of Low-Pass Filtering", <http://lands.let.kun.nl>
- [10] Iseli, M. R. and Alwan, A., "Inter- and Intra-speaker Variability of Glottal Flow Derivative Using the LF Model", http://citeseer.nj.nec.com/cache/papers/cs/17821/http:zSzwww.icsl.ucla.edu:zSz~spaplzSzpublicationszSziseli_ic_slp2000.pdf/iseli00inter.pdf
- [11] Boštík, M., *Analyza hlasu pro diagnostické účely*, Diploma work, BUT, Brno, 2002.