EXCITATION-SYNCHRONOUS GLOTTIS INVERSE-FILTERING BY MEANS OF A SELF-EXCITED THRESHOLD AUTO-REGRESSIVE MODEL OF THE SPEECH SIGNAL

*J. Schoentgen * and Z. Azami †*
Institute of Modern Languages and Phonetics , CP 110, Université Libre de Bruxelles, 50 Avenue F.-D. Roosevelt, 1050 Bruxelles, Belgium.
** National Fund for Scientific Research, Belgium, † Grant, U.L.B.*

ABSTRACT

We propose to carry out excitation-synchronous glottis inverse filtering by means of a compound auto-regressive model of the speech signal. The model consists of two linear auto-regressive models that are excitation-synchronously fitted by means of an auxiliary signal which has the same period as the speech signal and a single peak per cycle. Auxiliary and speech signals need not be aligned.

INTRODUCTION

Glottal inverse filtering is the estimation of the glottal waveform from the speech signal. Generally speaking, glottal inverse filtering consists of filtering the speech signal by means of an "inverted" transfer function estimate in which poles are replaced by zeros. Difficulties with glottal inverse filtering are the following. a) The vocal tract transfer function must be estimated from the speech signal. But, the speech signal is not the impulse response of the vocal tract. It is, in a first approximation, the convolution of the impulse response and the glottis signal which includes effects of the interaction between voice source and vocal tract. b) During connected speech, the transfer function varies with time whereas conventional estimation techniques (e.g. linear predictive analysis) posit that the speech signal is stationary during the analysis interval. c) Even during the emission of sustained vowels, the vocal tract transfer function cannot be assumed to be stationary because the vibrating vocal folds rhythmically connect and disconnect subglottal and tract cavities. As a result, both eigenfrequencies and bandwidths change within a glottal cycle. As a consequence, glottal signals are difficult to estimate reliably and, more often than not, attempts at inverse filtering are confined to sustained vowels.

Here, we propose to carry out inverse filtering in the following way. First, closed phases of the glottal cycle are detected via a compound speech signal model that consists of two linear auto-regressive models. Fitting of the model is carried out by means of the overall prediction error and the energy difference between the "open" and "closed" phase components of the speech signal. Second, formant frequencies and bandwidths of the "closed" phase components are estimated. Indeed, a conventional solution of problem (c) is to estimate the tract transfer function throughout the closed phases of the glottal cycles [1]. A consequence is that effects of the interaction between vocal tract and glottal source are included in the voice source signal [2]. Third, after eliminating real poles and complex poles whose bandwidths are larger than 500 Hz, inversed filtering is carried out by means of a cascade of second-order cells, that is one cell per pair of complex conjugate poles.

MODEL

The compound model of the speech signal we have proposed earlier is the following [3] [4] [5] [6].

$$y(n) = a_0 + \sum_{i=1}^{N} a_i y(n-i), \quad w(n-d) < r \tag{1}$$

$$y(n) = b_0 + \sum_{i=1}^{M} b_i y(n-i), \quad w(n-d) \geq r \tag{2}$$

Signal $y(n)$ is represented by means of two linear auto-regressive models (1) & (2). n is the time index, r a threshold and d a delay. a_i and b_i are the coefficients of linear sub-models (1) & (2) and N and M their orders. $w(n)$ is an auxiliary signal that must have a single vertex per cycle and the same period as signal $y(n)$. However, signals $w(n)$ and $y(n)$ need not be aligned.

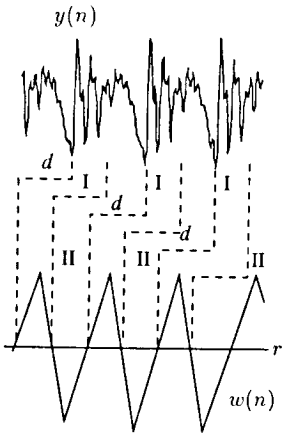

Figure 1: Speech signal $y(n)$ and synthetic triangular auxiliary signal $w(n)$.

METHOD

In the framework of glottal inverse filtering, we made use of two synthetic auxiliary signals. The first was a rectangular and the second an isosceles triangular waveform. The difference between rectangular and triangular auxiliary signals was that for the latter the length of cutouts (I) could be varied between 0 (the summit of the triangle) and period T (its base) by means of threshold r. The function of delay d was to position these lengths, cut out by means of threshold r, with respect to the time coordinates of the intersections of threshold r with signal $w(n)$ (Fig.1). Delay d and threshold r were either determined by means of an optimizer or a systematic search. When auxiliary signal $w(n)$ was a rectangular waveform, the cutout lengths were constants equal to the crenellation width and the only variable whose optimal value had to be searched for was delay d. The crenellation width of rectangular auxiliary signal $w(n)$ was fixed at 40 % of glottis cycle length T. Therefore, the lengths of cutouts (I & II) were respectively equal to $w(n)$'s crenellation width and the remaining 60 % of the fundamental period. Delay d was varied between 0 and T. For each choice of delay d, sub-models (1) & (2) were fitted to their respective cutouts (I) & (II) by means of singular value decomposition and the normalized overall prediction error was calculated. Sub-model orders were respectively equal to 9 and 8 and the sampling frequency was equal to 8 kHz. The best break-up of the speech signal into "open" and "closed" phase cutouts was the one that gave rise to a local minimum of the prediction error and to the biggest difference between the energies of signal components (I) & (II). In other words, the "closed" phase components of the speech signal were assigned to those cutouts that gave rise to a local minimum of the overall prediction error and to a maximum of the signal energy. For this choice of d, formant frequencies and bandwidths of the "high-energy" cutouts were determined and inverse filtering was carried out by means of a cascade of second-order cells.

The stages of the segmenting, fitting and inverse filtering algorithm were as follows :

(i) Asynchronous positioning of an analysis window of a length of 26 msec ;
(ii) Estimation of glottis cycle length T by means of the smoothed speech signal ;
(iii) Initialization of delay d, d=0 ;
(iv) Segmentation of the windowed speech signal (cf. Figure 1) by means of a rectangular auxiliary signal ;
(v) Least mean square fitting of sub-model (1) to cutouts (I) and of sub-model (2) to cutouts (II) ;
(vi) Calculation of the total normalized prediction error. Normalization was by the cutout lengths ;
(vii) Incrementation of d. If $d <$ T then step (iv) otherwise step (viii) ;
(viii) Calculation of the energy differences between components (I) and (II) for delays d that gave rise to the five smallest prediction errors ;
(ix) Selection of the segmentation (on the base of delay d) that gave rise to a maximum energy difference during step (viii) ;
(x) Reestimation by means of a covariance multi-interval method of linear predictive coefficients a_i of components (I) arrived at step (ix) ;
(xi) Computation of formant frequencies and bandwidths by means of the predictive coefficient polynomial ;
(xii) Discarding of real poles and of complex pole pairs giving rise to bandwidths larger than 500 Hz ;
(xiii) Inverse filtering by means of a cascade of second-order cells, i.e. a cell per pair of complex conjugate poles remaining after step (xii) ;
(xiv) Segmentation by means of period T and delay d of the glottal waveform so arrived at ;

RESULTS AND DISCUSSION

Figures 2 and 3 show glottal waveforms, of sustained vowels or vowel transitions, obtained by means of the inverse filtering method previously explained. The displayed waveforms of four speakers were arrived at wholly automatically. It is seen that they have traits that are typical of waveforms that have been obtained in the framework of other studies. Part of the observed inter- and intra-speaker variability is generally believed to be a consequence of the fickle ability of the linear predictive model to represent the speech signal adequately. But, variability may also have been a consequence of the occasional inability of steps (iii) to (viii) to segment identically from one analysis window to the next. Also, we have tried out, on the same speech signals, auxiliary signals of rectangular and triangular shape. The ability of the triangular waveform to give rise to variable cutout lengths did however not appear to be an advantage over the rectangular waveform whose cutout lengths were fixed.

It is planned to post-process waveforms so as to get rid of pulse estimates that are outliers and handle intra-speaker variability via vector quantization which chooses a representative set of glottal pulses. Indeed, here the purpose of inverse filtering is not to provide entire glottal waveforms. Instead, the objective is to arrive at a set of speaker-typical glottal pulses so that the discrimination performance of acoustic features, related either to the glottal pulse or the vocal tract transfer function, can be compared in the framework of a speaker recognition task. The purpose is to find an answer to the question of whether speaker identity is made up of acoustic cues that bear on the voice source, or the vocal tract or a combination of the two.

References

[1] A. K. Krishnamurthy and D. G. Childers. Two-channel speech analysis. *IEEE Trans. Acoust., Speech and Signal Processing*, ASSP-34(4):730-743, 1986.

[2] T.V. Ananthapadmanabha and G. Fant. Calculation of true glottal flow and its components. *Speech Communication*, 1:167-184, 1982.

[3] J. Schoentgen and Z. Azami. Pitch-synchronous formant extraction by means of a compound auto-regressive model. *Proceedings Eurospeech'93*, pages 401-404, 1993.

[4] Z. Azami and J. Schoentgen. Extraction des formants en synchronie avec l'excitation à l'aide d'un modèle auto-régressif composé. *Proceedings XXèmes Journées d'Etude sur la Parole*, pages 235-240, Trégastel, 1994.

[5] J. Schoentgen. Self-excited threshold auto-regressive models of the glottal pulse and the speech signal. *Proceedings of the ICSLP*, pages 1063-1066, 1994.

[6] J. Schoentgen. Dynamic models of the glottal pulse. In *Levels in Speech Communication, Relations and interactions, Sorin et al. (Eds), Elsevier, Amsterdam*, pages 249-266, 1995.

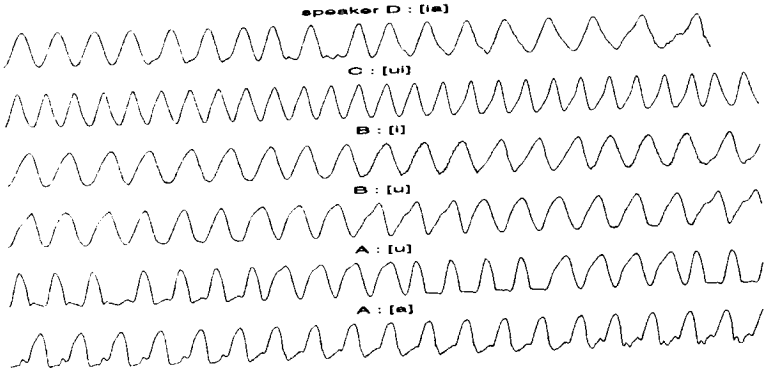

Figure 2:

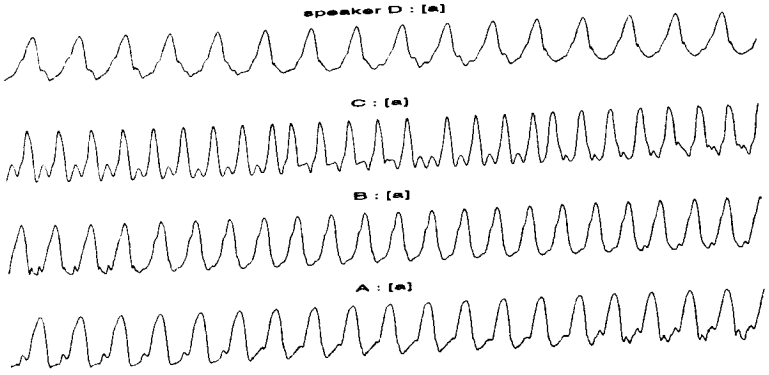

Figure 3: