A ROBUST METHOD FOR DETERMINING INSTANTS OF MAJOR EXCITATIONS IN VOICED SPEECH

B. Yegnanarayana
Indian Institute of Technology, Madras-600036, India

R.L.H.M. Smits
Institute for Perception Research, 5600 MB Eindhoven, The Netherlands

ABSTRACT

In this paper we propose a method for determining the instants of significant excitation in speech signals. Here significant excitation refers primarily to the instants of glottal closure in voiced speech. The method computes the average slope of the unwrapped phase spectrum as a function of time. The instants where the phase slope function makes a positive zero-crossing correspond to the major excitations in the signal. For an analysis window size in the range of one to two pitch periods, these instants coincide with the instants of glottal closure in each pitch period. The method is robust, as it depends only on the average phase slope value, and further, it dpends only on the positive zero-crossing instants of the average phase slope function.

1. INTRODUCTION

Voiced speech is produced as a result of excitation of the vocal tract system by a quasiperiodic sequence of glottal pulses. Within each period the major excitation takes place at the instant of glottal closure. We call these instants *significant* instants. In this paper we propose a method of determining these instants of significant excitation automatically from a speech signal using the negative derivative of the unwrapped phase (group delay) function of the short time Fourier transform of the signal [1].

Many speech analysis situations depend on the accurate estimation of the instant of glottal closure within a pitch period. For example, if such instances are known, the closed glottis region can be identified, and the vocal tract parameters such as formants may be derived accurately by confining the analysis to only those regions [2]. It is also possible to determine the characteristics of the voice source by a careful analysis of the signal, starting with this information [3].

Several methods have been proposed for determining the instant of glottal closure [2,3,4]. Almost all of them use some kind of block processing to determine the energy of the residual excitation signal in a small interval. The point where the computed energy is maximum is marked as the instant of significant excitation. While these methods work well in most cases, the block processing leaves some uncertainty as to the precise location of the instant of excitation [3,4].

In this paper we present a method for determining the instants of significant excitation using the properties of minimum phase signals and group delay functions [1]. In Section 2 we discuss the basis for the proposed method. Development of the algorithm for determining the significant instants of excitation is described in Section 3. In Section 4 we consider an example of natural speech data to demonstrate the applicability of the method.

2. BASIS FOR THE PROPOSED METHOD

Consider a delayed (τ) unit sample sequence shown in Fig.1(a). The Fourier transform (FT) of the sequence is $\exp(-j\omega\tau)$. The FT phase function is $\phi(\omega) = -\omega\tau$, and its negative derivative (or group delay), shown in Fig.1(b), is $-\phi'(\omega) = \tau$. Thus the phase function has a constant slope which corresponds to the delay of the unit sample in the time domain. Let us assume an analysis window enclosing the unit sample. As the window is moved to the right or left, the delay of the unit sample changes with respect to the position of the window. The average value of the negative derivative of the FT phase (group delay function) varies linearly with the position of the analysis window. The average value of the group delay as a function of the position of the window is called phase slope function. The instant at which the phase slope function crosses zero is identified as the delay of

0-7803-2431-5/95 $4.00 © 1995 IEEE

the unit sample in the time domain.

Now consider a delayed damped sinusoid as shown in Fig.1(c). The negative derivative of the FT phase for the signal is shown in fig.1(d). The average value of the derivative (phase slope) is equal to the delay of the damped sinusoid with respect to the position of the window. As the analysis window is moved, again the phase slope value varies linearly with time.

In general a minimum phase signal starting at time $t = 0$ has the property that its average value of the unwrapped FT phase spectrum is zero. If the signal is delayed, then the average slope of the phase spectrum is proportional to this delay. This is the basis for the proposed method of deriving the instants of significant excitation.

3. ALGORITHM FOR EXTRACTING SIGNIFICANT INSTANTS

In this section we describe the development of an algorithm for determining the instants of significant excitation from the average slope of the phase spectrum of the signal. First we discuss the computation of the average slope of the phase spectrum, and then we discuss a method to extract the desired instants from the phase slope values as a function of time. The main step is the computation of average slope of the unwrapped phase spectrum. Direct computation of the phase spectrum through the real and imaginary parts of the the DFT of a signal results in phase values which are confined to the range $-\pi$ to $+\pi$. In other words, the phase values are said to be wrapped around these limits. To compute the average slope, it is necessary to determine the unwrapped phase values. Accuracy of computation of the unwrapped phase depends on the (windowed) signal and the true phase values. Since we need only the average value of the group delay function, we can avoid computation of the unwrapped FT phase by computing the group delay function directly from the windowed signal $x(n)$. If $X(\omega)$ and $Y(\omega)$ are the FTs of $x(n)$ and $nx(n)$, respectively, then the group delay is given by [5]

$$-\phi'(\omega) = \tau(\omega) = (X_R Y_R + X_I Y_I)/(X_R^2 + X_I^2)$$

where $X_R + jX_I = X(\omega)$ and $Y_R + jY_I = Y(\omega)$. Isolated peaks in $\tau(\omega)$ are removed by using a 3-point median smoothing. The average value of the smoothed $\tau(\omega)$ is computed. The resulting phase slope function is computed by moving the analysis window by one sample at a time. Figs.2(c) and 2(d) show the phase slope functions for the output (shown in Fig.2(b)) of an all-pole model excited by a periodic impulses sequence (shown in Fig.2(a)) for two different window sizes, namely, 100 and 50 samples, respectively. The positive zero-crossing instants of the phase slope function correspond to the instants of significant excitation. It can be seen from Fig.2(d) that, using smaller windows it is possible to identify minor excitations as well. Since the group delay function is affected by truncation due to windowing, the phase slope function may deviate randomly from the ideal straight line, as can be seen Figs.2(c) and 2(d). However, the zero crossing instants are not significantly affected by these random fluctuations.

4. EXTRACTION OF INSTANTS FROM SPEECH SIGNAL

Due to discrete nature of computations in obtaining the phase slope function and also due to effects of finite window size and shape, the phase slope function for real speech data will show many fluctuations which may sometimes make it difficult to identify the positive zero-crossing instants uniquely. Figs.3(a) and 3(b) show a segment of voiced speech and its linear prediction (10th order) residual[6]. Fig.3(c) shows the phase slope function computed for the residual signal shown in Fig.3(b). The choice of the residual signal, instead of speech signal directly, reduces the truncation effects of windowing without altering the information about the significant instants. Note that the inverse filter used to obtain the LP residual is also a minimum phase filter, and hence the information about the instants of significant excitation is not altered in the residual signal.

Due to approximation in the computation of the average value, the phase slope function is not a smooth straight line. It can be smoothed using a linear smoothing filter like mean filtering. The parameters of the mean filter are not very critical as the positive zero-crossing instants do not change significantly due to this filtering. Fig.3(d) shows the extracted instants information.

5. CONCLUSIONS

In this paper we have proposed a method to determine the instants of significant excitation using the average group delay characteristics of minimum phase signals. Since the method is based on the phase characteristics of the excitation signal, it is possible to derive the instants of significant excitation for all categories of speech segments. The method works well even for female voices because there is no influence of the vocal tract system on the phase slope characteristics of the signal. The method is also not very sensitive to the choice of analysis parameters, like the size of the window and the placement of the window relative to the significant excitation instants. It is interesting to note that by block processing we have been able to mark an instant that does not depend crit-

ically on the size of the block and its placement.

References

[1] Yegnanarayana, B. (1984) Significance of group delay functions in signal reconstruction from spectral magnitude or phase. *IEEE Trans Acoust., Speech and Signal Processing 32(4)*, 610–623.

[2] Krishnamurthy, A. K. (1992) Glottal source estimation using a sum of exponentials model. *IEEE Trans. Acoust., Speech and Signal Processing 40(3)*, 682–686.

[3] Ananthapadmanabha, T. V. and Yegnanarayana, B. (1979) Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Trans. Acoust., Speech and Signal Processing 27(8)*, 309–319.

[4] Wong, D.J., Markel, J.D. and Gray, A.H. (1979) Least squares glottal inverse filtering from the acoustic speech wave. *IEEE Trans. Acoust., Speech and Signal Processing 27(8)*, 350–355.

[5] Oppenheim, A.V. and Schafer, R.W. (1975) *Digital Signal Processing. ch.10.* Prentice-Hall, Englewood Cliffs, NJ.

[6] Markel, J.D. and Gray, A.H. (1976) *Linear Prediction of Speech.* New York: Springer-Verlag.

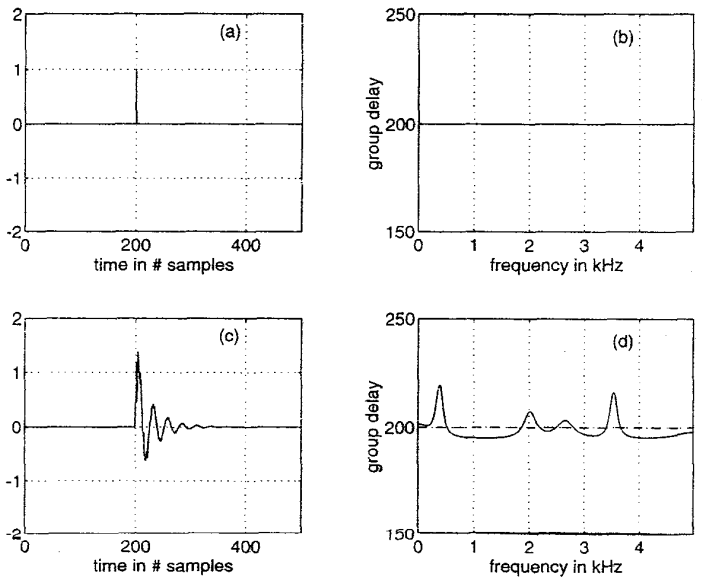

Figure 1. (a) Delayed unit sample sequence. (b) Group delay function for delayed unit sample sequence in (a). (c) Response of an all-pole filter for the unit sample sequence in (a). (d) Group delay function for the signal in (c).

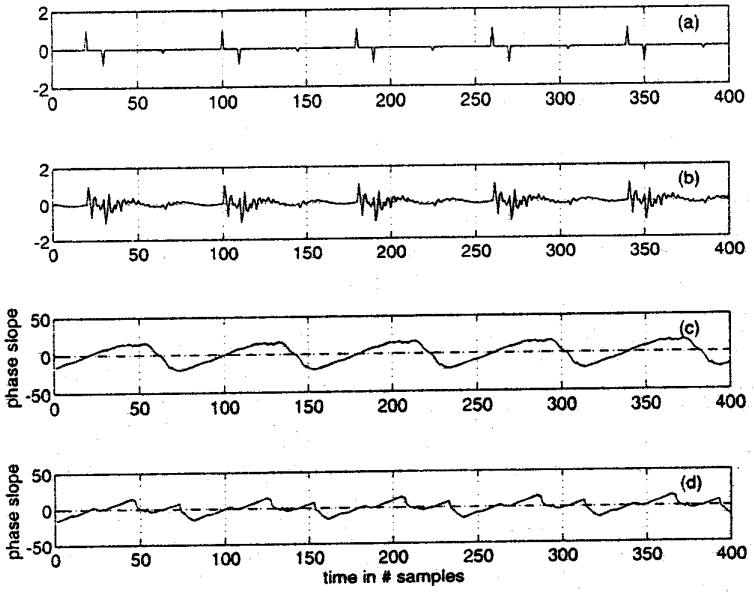

Figure 2. (a) Periodic excitation impulses sequence. (b) Response of an all-pole filter to the impulses sequence of (a). (c) Phase slope function for the signal in (a), for an analysis window of size 100 samples. (d) Phase slope function for the signal in (a), for an analysis window of size 50 samples.

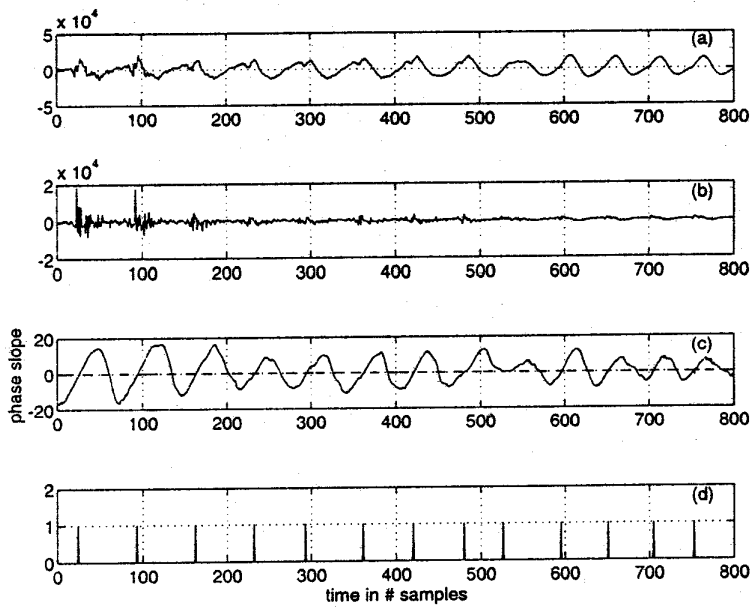

Figure 3. (a) A segment of a speech waveform. (b) Linear prediction (10th order) residual of the speech segment in (a). (c) Phase slope function of the signal in (b) using an analysis window of 100 samples. (d) Positive zero-crossing instants of the signal in (c), which correspond to instants of significant excitation.