On timing in time-frequency analysis of speech signals

B YEGNANARAYANA

Department of Computer Science and Engineering, Indian Institute of Technology, Madras 600 036, India

Abstract. The objective of this paper is to demonstrate the importance of position of the analysis time window in time-frequency analysis of speech signals. Speech signals contain information about the time varying characteristics of the excitation source and the vocal tract system. Resolution in both the temporal and spectral domains is essential for extracting the source and system characteristics from speech signals. It is not only the resolution, as determined by the analysis window in the time domain, but also the position of the window with respect to the production characteristics that is important for accurate analysis of speech signals. In this context, we propose an event-based approach for speech signals. We define the occurrence of events at the instants corresponding to significant excitation of the vocal tract system. Knowledge of these instants enable us to place the analysis window suitably for extracting the characteristics of the excitation source and the vocal tract system even from short segments of data. We present a method of extracting the instants of significant excitation from speech signals. We show that with the knowledge of these instants it is possible to perform prosodic manipulation of speech and also an accurate analysis of speech for extracting the source and system characteristics.

Keywords. Time-frequency analysis; group delay; glottal vibration; prosodic manipulation; voice source; vocal tract system.

1. Introduction

Many natural signals contain useful spectral information over a range of frequencies, and these spectral features may vary with time. In order to perform spectral analysis a windowed time domain signal is chosen. The size of the window is dictated by the desired resolution in the frequency domain. The shape of the window is dictated by the edge effects due to abrupt termination of the signal. Effect of windows on time-frequency analysis of signals has been well studied in the literature (Harris 1978). Usually shorter windows provide better temporal resolution but are suitable only for high frequency signals. On the other hand, longer windows provide better frequency resolution but mask features due to fast temporal variations. Several methods have been proposed addressing the issues of adapting the effective window size for achieving temporal and spectral resolutions simultaneously (Hlawatsch & Boudreaux-Bartels 1992). Most of the time-frequency analysis

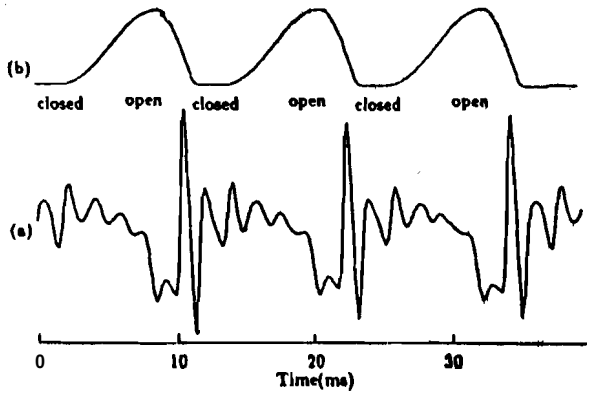

Figure 1. **Speech** signal and the corresponding glottal waveform showing the open and closed regions. **(a) Speech** waveform. (b) Glottal waveform.

methods consider mainly the issue of resolution. But the positioning of the window relative to the signal has significant influence on the results of analysis (Rabiner ***et al*** 1977). For example, if the analysis window contains transition from one steady state to another, it is difficult to associate a steady system with the derived spectrum. Moreover, in order to examine the variability of the system over a period of time, it is necessary that the system characteristics are extracted from steady portions in the region. That means that the analysis window has to be positioned suitably with respect to the signal. We call this positioning of window as ***timing*** in ***time-frequency analysis.***

The timing becomes crucial for the time-frequency analysis of speech signals. The objective in speech analysis is to extract the characteristics of the excitation source and the vocal tract system from speech signals. Since both the source and the system vary during production of speech, it is necessary to use short windows for analysis. Note that even within a pitch period of voiced speech, the source is not steady due to glottal vibration. Also the vocal tract system is not steady due to interaction between source and system. Therefore even pitch synchronous analysis of voiced speech does not overcome the problem of position effects of the analysis window mentioned above. The pitch synchronous placement of window does not guarantee that the signal within the window corresponds to a steady system (Parthasarathy & Tufts 1987).

In voiced speech the vocal tract system is excited by a periodic glottal vibration. The glottal vibration is due to slow opening and sudden closing of the vocal folds, followed by a closed phase during each pitch period (see figure 1). Significant excitation of the vocal tract system takes place during the rapid closing part of the glottal vibration. The vocal tract system characteristics are preserved in the signal in the closed glottis region. The vocal tract resonances are damped significantly during the opening phase of the glottal vibration, when the trachea is coupled to the vocal tract system. Thus the vocal tract system characteristics are significantly different during the open and closed phases of the glottal vibration. In addition, the excitation source will have a turbulent noise component due to air passing through the narrow constriction created during the closing phase of the glottal vibration. Thus a straightforward analysis of the speech signal, even using a pitch synchronous window, is not likely to bring out the time varying source and system characteristics.

In order to represent the time varying vocal tract system characteristics across several pitch periods, it is necessary to use an analysis window enclosing the steady vocal tract system region in each pitch period, and compare the system characteristics derived from similar regions for all pitch periods. These are two steady regions in each pitch period, one corresponding to the closed glottis region, and the other to the open glottis region.

Thus it is necessary to identify these regions first and then select a suitable window within each region for analysis. One way to do this is to determine first the instant of significant excitation in each pitch period. Then a short (2-3 ms) segment to the right of the instant can be considered the closed phase region and a short (2-3 ms) segment to be left of the instant the open phase region. Even though this division provides only an approximation to the selection of the regions, analysis of the vocal tract system corresponding to, say the closed phase region, across successive pitch periods will definitely bring out the time varying vocal tract resonance characteristics accurately. Moreover, the knowledge of the instants of significant excitation also permits an accurate analysis of the source characteristics.

We call these instants of significant excitation of the vocal tract system as events, and the analysis based on the knowledge of these events as event-based analysis. It is obvious that event-based analysis will provide better results than the results from arbitrary placement of windows for analysis. But it is necessary to determine these events first before performing time-frequency analysis.

In this paper we describe a method for extracting the instants of significant excitation from speech signals, and show how the knowledge of these instants can be used for various applications in speech processing. In §2 we discuss the signal processing principles used for extraction of the instants of significant excitation. In §3 we describe a method of extraction of these instants from speech signals. We show in 94 how the knowledge of these extracted instants can be used for prosodic manipulation, which involves modifying the pitch periods and durations of various speech segments in a predetermined manner without modifying the vocal tract system characteristics. In §5 we give some results of analysis of the source characteristics by using two successive pitch periods at a time. In 96 we demonstrate the use of the instants of significant excitation for extracting the time varying characteristics of the vocal tract system in the closed and open glottis regions separately from successive pitch periods.

2. Principle of the proposed method of extraction of instants of significant excitation (Yegnanarayana & Smits 1995)

Consider a unit sample sequence with the sample at t $= \tau$ as shown in figure 2a. Then its Fourier transform (FT) is $e^{-j\omega\tau}$ and hence the FT phase is given by $\phi(\omega) = -\omega\tau$. Thus the derivative of phase or the negative group delay is given by $\phi'(\omega) = -\tau$. As the analysis window, centred around the origin t = 0 and enclosing the unit sample, is moved to the right, $\phi'(\omega)$ will vary linearly with time, crossing through zero at t $= \tau$, as shown in figure 2d. The plot in figure 2d is called phase slope function. If we consider a unit sample response (shown in figure 3a) of a second order all-pole system corresponding to a damped resonator, the average of $\phi'(\omega)$ will be equal to the delay of the unit sample from the origin. By moving the analysis window to the right, the average $\phi'(\omega)$ will increase linearly with time passing through zero at t $= \tau$, as shown by the phase slope function plot in figure 3c. Thus for each analysis window position it is necessary to compute the group delay and to find the average of the group delay spectrum. The group delay spectrum is computed as follows.

Let $x(n)$, n $= 0, 1, \ldots$ N $-$ 1 be the given signal in the analysis window and let $y(n) = nx(n)$, n $= 0, 1 \ldots$ N $-$ 1. Then the derivative of the FT phase or negative group

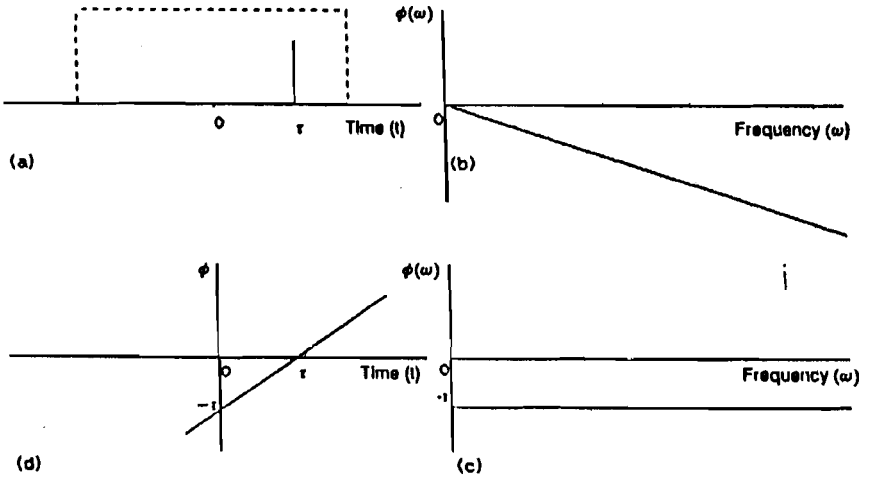

Figure 2. Illustration of group delay properties of a unit sample. **(a)** Unit sample sequence. (b) FT phase of unit sample sequence. **(c)** Derivative of FT phase. (d) Phase slope function.

delay is given by (Oppenheim & Schafer *1989,* ch. *12).*

$$\phi'(\omega) = -[X_I(\omega)Y_I(\omega) + X_R(\omega)Y_R(\omega)]/[X_I^2(\omega) + X_R^2(\omega)] \tag{1}$$

where the FT of $x(n)$ is given by $X(\omega) = X_R(w) + jX_I(w)$ and the FT of $y(n)$ is given by $Y(w) = Y_R(w) + jY_I(w)$. The derivative of the FT phase $\phi'(\omega)$ is computed from the discrete Fourier transforms of $x(n)$ and $y(n)$ using a suitable order for DFT, preferably greater than twice the size of the analysis window.

Since $\phi'(\omega)$ is computed using a windowed signal, and also since it is available only at discrete frequencies, there will be large fluctuations in the computed values of $\phi'(\omega)$. To obtain the average value from the discrete values of $\phi'(\omega)$, the function is smoothed using a 3-point median filtering to eliminate large fluctuations, and then the mean value of the smoothed $\phi'(\omega)$ is computed. Due to computational inaccuracies, there may be some error

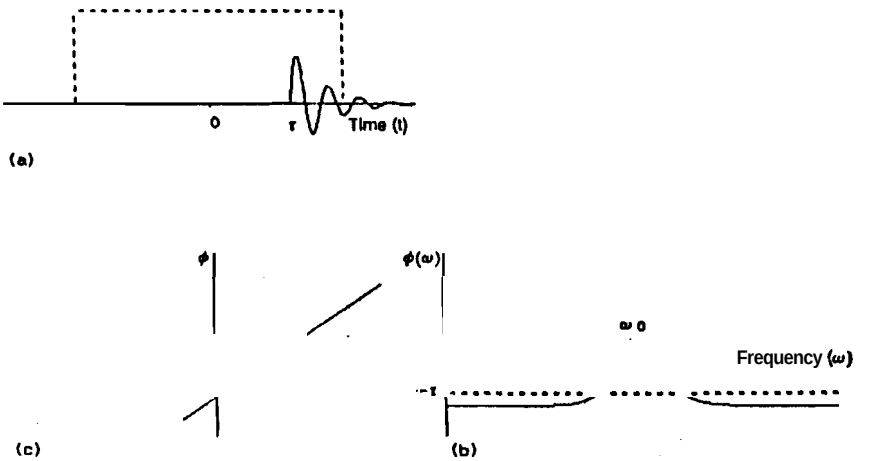

Figure 3. Illustration of group delay properties of a damped sinusoidal signal. **(a)** Delayed damped sinusoidal signal. (b) Derivative of the FT phase. **(c)** Phase slope function.

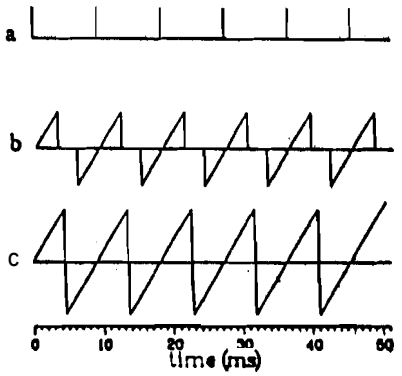

Figure 4. Phase slope function for a periodic impulse sequence. **(a)** Periodic impulse sequence. **(b)** Phase slope function for a window size of 6.4ms. **(c)** Phase slope function for a window size of 12.8ms.

in the resulting estimate of the mean value of $\phi'(\omega)$. The mean value $\bar{\phi}'$ is estimated at each instant by shifting the analysis window by one sample at a time, to obtain $\bar{\phi}'$ for that instant. We call the resulting function $\bar{\phi}'$ vs time the phase slope function. The positive zero crossing instant of the phase slope function gives the instant of significant excitation in the analysis window. It is also interesting to note that the phase slope function is not dependent on the phase characteristics of the system, as long as the system is a minimum phase system. This is because a minimum phase system excited by an impulse at $t = 0$ has zero average phase slope value (Berkhout 1974; Oppenheim & Schafer 1989, ch. 5).

3. Illustration of the method for synthetic and natural speech signals (Smits & Yegnanarayana 1994)

3.1 *Instants of excitation in synthetic signals*

The phase slope function is computed for several cases of synthetic signals as shown in figures 4–9. Examples are given for two different window sizes in each case. Each figure contains three plots: (a) the original time signal, (b) the phase slope function for an analysis window of size 6.4ms, and (c) the phase slope function for an analysis window of size 12.8ms. Figure 4 is for a periodic impulse sequence, figure 5 is for a periodic impulses sequence, and figure 6 is for the output of an all-pole model excited by the periodic impulses sequence. In figures 4-6 we can see that the positive zero-crossing instants of the phase slope functions correspond to the instants of major excitation within the chosen analysis window. In figure 4 a large portion of the phase slope function is linear around the instant

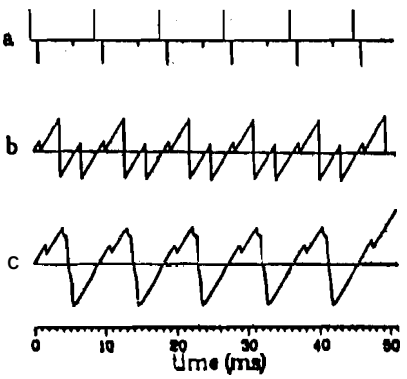

Figure 5. Phase slope function for a periodic impulses sequence. **(a)** Periodic impulses sequence. **(b)** Phase slope function for a window size of 6.4ms. **(c)** Phase slope function for a window size of 12.8ms.

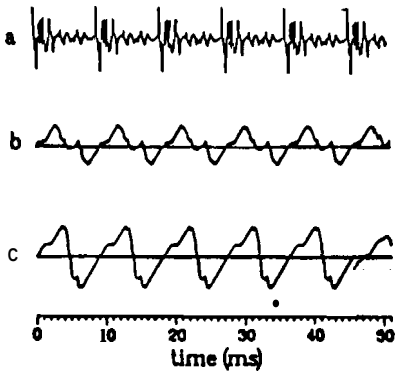

Figure 6. Phase slope function for the output of an all-pole model excited by the periodic impulses sequence of figure 5a. (a) Output of an all-pole model excited by a periodic impulses sequence. (b) Phase slope function for a window size of 6.4ms. (c) Phase slope function for a window size of 12.8ms.

of major excitation. This linear part is the result of dominance of one excitation in the analysis window.

If a signal containing minor excitations is analysed using a large-size window (figures 5 and 6), the phase characteristics due to major excitation dominate the phase slope function, and the minor excitations do not influence the positive zero-crossings. However, the presence of these minor excitations may sometimes make it difficult to identify the instants of major excitations.

Figures 4–6 also show that the phase slope function is mostly dictated by the excitation signal. It is interesting to note that neither the minimum phase all-pole system, nor the location and size of the analysis window has influenced the decision on the excitation instants obtained from the phase slope function. Even for a window size greater than a pitch period, the use of the Hanning window reduces the effects of the surrounding impulses on the resulting extraction of the excitation instants. This is because within a window only one major excitation impulse is likely to dominate.

For random noise (figure 7), the features in the phase slope function are different for different window sizes. It is interesting to note that any major excitation in the noise signal will clearly show up irrespective of the size of the analysis window. For any noise signal, these excitation instants will be distributed randomly in time. That is why the method will not work well for noisy speech, as the excitation due to noise will show up randomly in between the instants of glottal closure. Figure 8 shows the behaviour of the phase slope function for two different window sizes for a signal generated by exciting an all-pole filter with random noise.

For sinusoids (figure 9), there are no isolated major points of excitation, and the phase slope function does not show the characteristic linear part around the positive zero-crossing

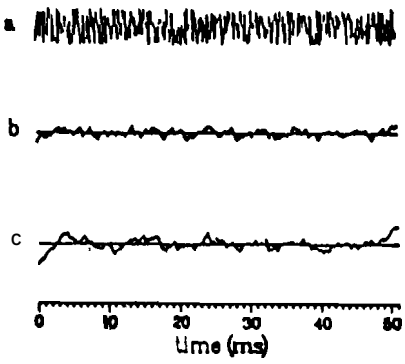

Figure 7. Phase slope function for a random noise sequence. (a) A random noise sequence. (b) Phase slope function for a window size of 6.4ms. (c) Phase slope function for a window size of 12.8ms.

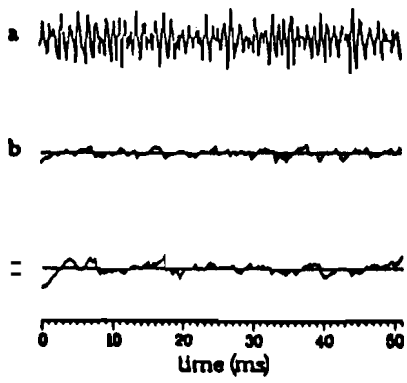

Figure 8. Phase slope function for an all-pole model excited by a random noise sequence. (a) Output of an all-pole model excited by a random noise sequence. (b) Phase slope function for a window size of 6.4ms. (c) Phase slope function for a window size of 12.8ms.

instants. The features of the phase slope function are different for different window sizes, and the effects of windowing show up clearly in the resulting phase slope function.

3.2 Instants of significant excitation in natural speech signals

In this section we discuss the performance of the proposed method on natural speech data. In all the illustrations to follow, the speech signals were sampled at 10 kHz. In order to minimize the effects of position of the analysis window with respect to the impulse response, it is better to obtain at least an approximation to the excitation signal before computing the average slope of the phase spectrum. Linear prediction (LP) residual (Markel & Gray 1976) is a good approximation to the excitation signal, as the correlation between adjacent samples is significantly reduced from what it is in the original signal. Note that since inverse filtering in linear prediction analysis is in effect passing the speech signal through a minimum phase system, the phase slope characteristics of the excitation will not be altered in the residual. For the computation of the residual, a 10th order linear predictor was used in this study, although the order is not very critical for this analysis.

Since for speech data there will be several points of excitation even within a pitch period, the phase slope function will have many fluctuations. In order to determine the instants of significant excitation, the points of positive zero-crossing of the phase slope function are obtained by smoothing the function.

Once the major excitations are identified, it is possible to explore for the presence of other excitation instants by computing the phase slope function on the residual using a smaller window for analysis, typically half the size of the original window. Thus the

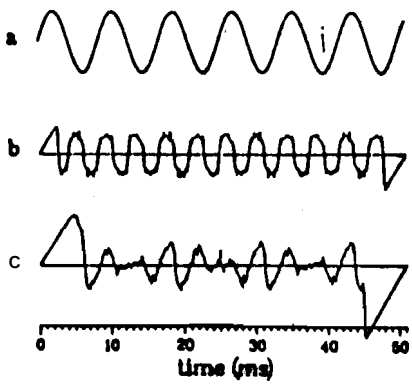

Figure 9. Phase slope function for a sinusoidal sequence. (a) Sinusoidal sequence. (b) Phase slope function for a window size of 6.4ms. (c) Phase slope function for a window size of 12.8ms.

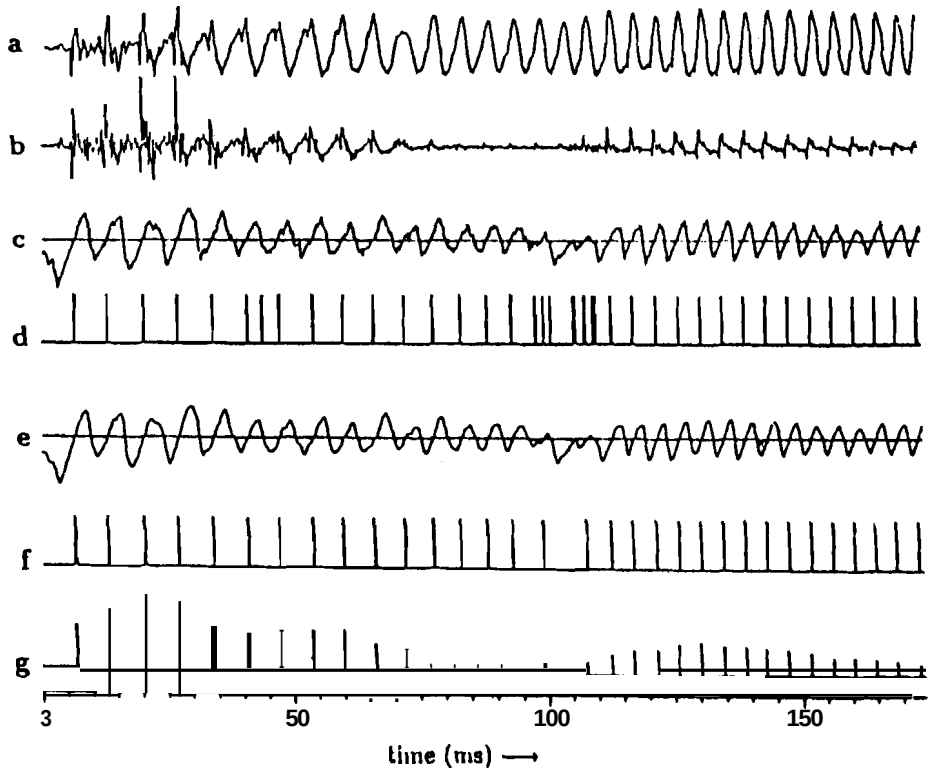

Figure 10. Illustration of extraction of the instants of significant excitation for voiced speech. (a) A segment of voiced speech signal. (b) Linear prediction residual of (a). (c) Phase slope function for the LP residual signal in (b). (d) Unit impulse sequence with impulses located at positive zero-crossing instants of the plot in (c). (e) Smoothed phase slope function. (f) Unit impulse sequence with impulses located at positive zero-crossing instants of the plot in (e). (g) Gain plot indicating the strengths of the impulses at the positive zero-crossing instants of (e).

method, in principle, enables us to determine other significant instants in the excitation. For identifying a single significant instant in each pitch period, it is preferable to have an analysis window size in the range of one to two pitch periods.

Figures 10a and b show a segment of voiced speech and its linear prediction residual, respectively. Figure 10c shows the phase slope function computed from the residual. Figure 10d gives all the positive zero-crossing instants in the phase slope function. To select the ones corresponding to significant excitation, the phase slope function is smoothed using a 13-point Hanning window. Note that the size of the smoothing window is not very critical, as long as it removes some fine fluctuations. The positive zero-crossing instant corresponding to the major excitation in the analysis window is not affected by this smoothing operation. The instants of positive zero-crossing in the smoothed phase slope functions are shown in figure 10f.

We have also computed the gain at each of the zero-crossing instants, by computing the square root of the average energy per sample in the LP residual for the interval between two successive zero-crossing instants. The interval is centered around the zero-crossing point under consideration. The resulting gain values are shown in figure 10g. These values may be viewed as strengths of the impulses at the selected instants.

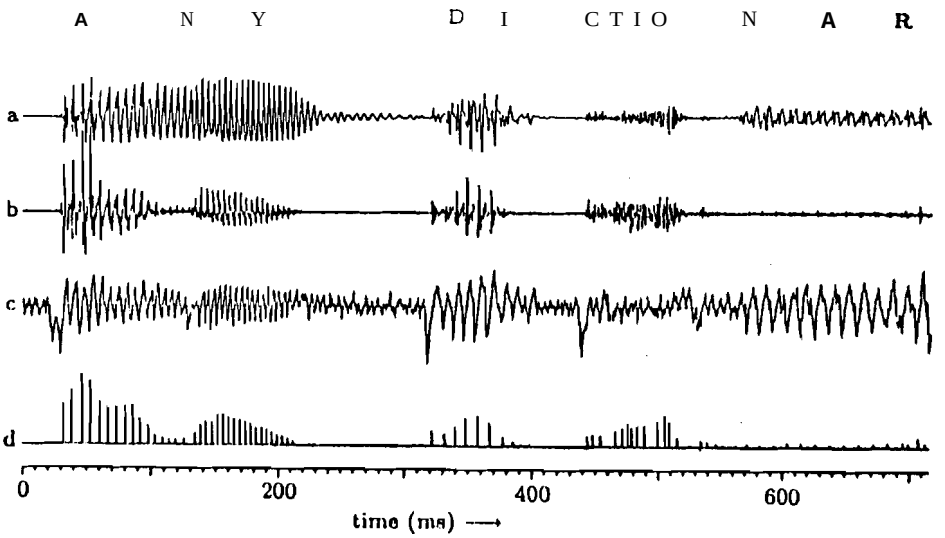

Figure 11. Illustration of extraction of the instants of significant excitation for a portion of the utterance for the sentence "ANY DICTIONAry will give..." by a male speaker. (a) Speech signal. **(b)** Linear prediction residual. (c) Phase slope function for the linear prediction residual signal. (d) Gain plot showing the strengths of the impulses at the significant instants.

3.3 Analysis of continuous speech

Figure 11 illustrates the result of our method on the initial part of the utterance "ANY DICTIONAry will give at least . . . ", uttered by a male voice. This utterance consists of a variety of segments like voiced, unvoiced, nasal, transition, stop, fricative etc. The length of the analysis window was 10ms, and the average pitch period was 8ms. Figures 11a and b show the speech waveform and the LP residual signal respectively. Figures 11c and d give the phase slope function and the gain plots, respectively, for the residual signal of figure 11b. The voiced segments have clear quasi-peperiodic instants of excitation. The unvoiced and silence parts show random instants of excitation. The instants (Figure 11d) identified by the phase slope function indeed correspond to the instants of significant excitation as can be seen by comparing with the residual signal in figure 11b.

Let us look at the individual signal categories in some detail. In the silence and unvoiced fricative regions, almost no significant excitation was identified. Note that in the unvoiced fricative regions the instants are at irregular intervals as in the case of noise, and hence are not significant. However, whenever there is a transition from one category to another, like at a burst release, the transition point is identified as a significant instant. In many weakly voiced regions there will not be any significant excitation, as evidenced by the residual signal. The same is reflected in the gain plot, although the speech waveform shows low frequency periodicity. Absence of significant excitation instants in this case can be verified by observing the amplitude spectra for these regions. Typically the spectra do not show any resonance or formant structure in these cases, but only show some energy at the pitch frequency.

Since the technique uses the phase characteristics of the excitation signal, the vocal tract system has very little influence on the proposed method of determining the instants of excitation. That is why it can be seen that the technique works well not only for steady vowels, but also for diphthongs, transitions, liquids and nasals. Note that the method shows

significant quasiperiodic excitations even in cases where they are not clearly evident in the linear prediction residual.

The technique works equally well for female voice speech also for all categories of segments. Note that because of the smaller average pitch period for female voices, a smaller analysis window size may be required.

4. Prosodic manipulation (Yegnanarayana & Teunen 1994)

In many applications and for studies in speech perception it is often desirable to generate speech with specified characteristics or to modify a given speech signal by incorporating some specified features. The features may include changes in the vocal tract system and source characteristics. These characteristics at a segmental level may correspond to, for example, the average pitch, vocal tract length and the source-tract interaction within each pitch period. At the suprasegmental level, the characteristics of interest are the durations of units at syllable or high levels, intonation and the speaking rate. Here we address the issue of modifying a given speech signal to incorporate specified features mainly at suprasegmental level. The emphasis is on the manipulation of prosodic features such as speaking rate and intonation (Moulines & Laroche 1995).

Availability of the instants of significant excitation makes prosodic manipulation easier, in principle, as it is these instants that need to be modified to realize any desired prosodic

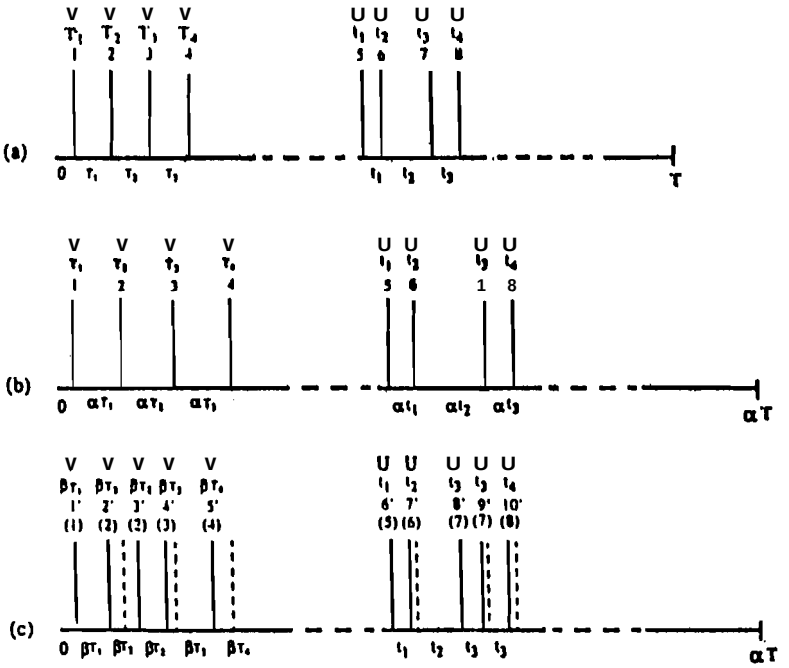

Figure 12. Illustration of prosodic manipulation. V and U are the voiced and unvoiced labels for the instants. $T_i's$ are intervals between instants in voiced regions, and $t_i's$ are intervals between instants in unvoiced regions. (a) Instants in the input data. (b) Instants shifted due to time scale multiplication factor α. (c) The new instants and the entries in the pitch period field at each instant in voiced and unvoiced regions, where the pitch period is modified by a factor @Note that the spacing between impulses is $\beta\ T_i$ in voiced regions and t_i in unvoiced regions.

characteristics. We focus mainly on the issue of manipulation of speaking rate and pitch period, although it is also possible to affect changes in the segmental characteristics as well. We discuss the procedure to incorporate the desired prosodic modifications. However, the procedure to derive the modification rules themselves is not within the scope of this work.

The data available for prosodic manipulation are the speech signal, the instants of significant excitation in the form of a gain function, and the linear prediction coefficient (LPCs) data file with Voiced (V)/ Unvoiced (U) labels. Centred around each of these instants a windowed speech signal is taken, and a residual signal is obtained by passing the speech signal through the inverse filter defined by the predetermined LPCs for the segment. From the residual signal around the instant, the required number of samples are taken to associate with the current instant. The gain per sample is computed at the instant by computing the square root of the mean squared energy of the residual signal associated with the instant.

The basic approach in prosodic manipulation is to derive an excitation signal incorporating the desired modification in the speaking rate and the pitch period. This is done by first taking the instants information in the gain function, and creating new instants data incorporating the speaking rates and pitch period modifications specified in the form of scale factors for these parameters. We associate with each instant, the time, pitch period (interval between successive instants), LP residual and LPCs. For speaking rate/duration manipulation, we obtain the new scaled time instants using the time scale manipulation factor. Likewise, for pitch manipulation, the pitch period associated with each instant is scaled appropriately. Now a new set of instants and the parameters at these new instants are determined as follows (see figure 12):

Proceeding from left to right, the first instant is copied as a new instant. The next new instant should be at the pitch period away from the first one, the period information being available in the parameter list associated with the first instant. Determine which of the old instants are closer to the new instant. Associate the parameter information of the old instant to the new instant. It is also possible to obtain an interpolated value of the pitch period for the current new instant from the pitch periods at the old instants which are on either side of the current new instant. Use the pitch period value in the parameter list at the current new instant to obtain the next new instant. This process is repeated until the end of all the instants derived from the original speech data.

Problems will arise while copying the residual samples at the new instants from the parameter list associated with the old instants, if the new pitch period is smaller than the old value at that instant. The required number of residual samples around the instant are copied. But to avoid discontinuity due to this partial selection of the residual samples, the residual signal samples are multiplied with a Hanning window. The signal is high pass filtered (cut-off frequency of about 50 Hz) to remove the very low frequency components including the zero frequency component. This will ensure that the resulting residual signal has zero mean. This process may produce some distortion, especially when the pitch period is scaled down significantly, say by a multiplication factor of 0.5 or lower. If the scaled pitch period is larger than the old one, the additional excitation samples needed in each pitch period are set to zero. The resulting excitation samples are appropriately scaled to obtain the gain value specified in the parameter list for the instant.

For instants labelled as unvoiced (U), the required number of residual samples are copied from the residual signal associated with the instants. For these instants, the entry in the pitch period field associated with the instants is not modified. Therefore if the interval between instants increase due to expansion of the time scale (slow speaking rate), some segments of the residual samples belonging to the unvoiced portion may be repeated. Sometimes this

will produce some audible distortion. One way to overcome this is to use random samples with appropriate gain, instead of repeating the residual samples.

Speech signal is generated by exciting the all-pole model, defined by the LPCs and the gain parameter, with the new excitation signal. It is also possible to vary the all-pole model characteristics within a pitch period to reflect the differences in the vocal tract system in the closed and open phases of the glottal vibration. This is realized approximately by using in the open phase a set of LPCs which correspond to the poles moved towards the origin in the z-plane. This creates an effect of damping of formants in the all-pole model representing the vocal tract system. The damping effect can be controlled by using a parameter to modify the LPCs. The parameter is simply the radius of a circle in the z-plane concentric with the unit circle.

It is also possible to generate the excitation signal using a model for the glottal pulse for voiced segments and random noise for unvoiced segments, and appropriately synchronizing them with the information associated with the instants. The glottal pulse model could be a model similar to the LF model described in the literature (Childers & Wong 1994).

5. Analysis of voice source characteristics

In this section we shall examine how the knowledge of the instants of significant excitation can be used for a careful analysis of source characteristics, especially for voiced speech. So far we have assumed only one major excitation within each pitch period, and hence we have used an analysis window of size about 1 to 2 pitch periods to extract the instants of significant excitation. These instants correspond to the instants of glottal closure. In addition, we may have an excitation, although minor, at the instant of glottal opening also. It is possible to extract these instants also, if we use an analysis window of size

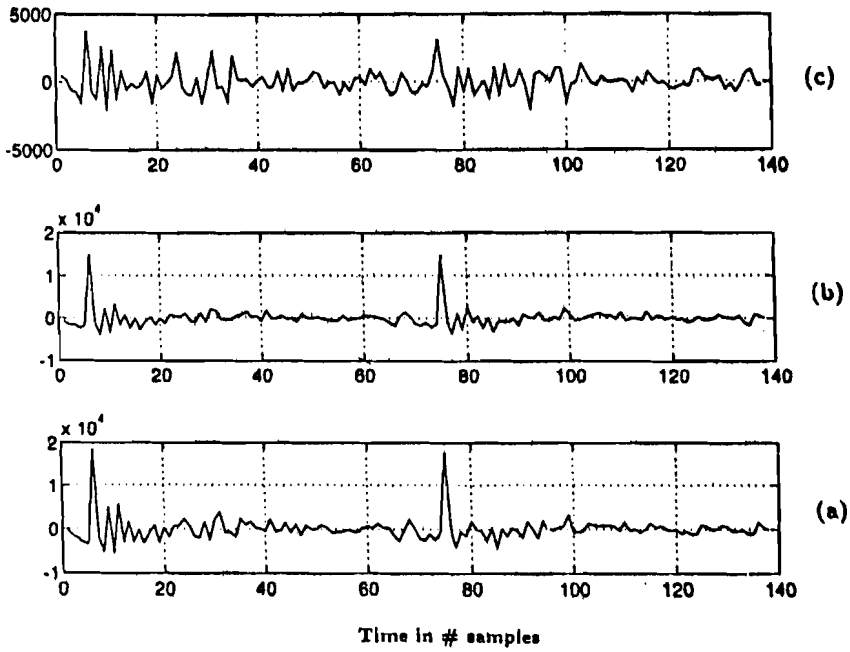

Figure 13. Decomposition of LP residual into deterministic and random parts. **(a)** Two periods of LP residual. **(b)** Deterministic part of the LP residual. (c) Random part of the LP residual.

less than one pitch period, typically about half the pitch period. Then we will have one additional positive zero-crossing of the phase slope function within each pitch period, and this corresponds to the opening instant. This can be seen in figure 5b, where the second positive zero-crossing corresponds to the instant of the small negative pulse. Note that it will not appear if the window size is of the order of a pitch period as in figure 5c. Thus if the excitation at the instant of glottal opening is also significant, then it can be extracted with the proposed method using a smaller window. These minor excitation instants can be identified after extracting the instants of major excitation at the glottal closure using a longer window.

The source for voiced speech also has a component of turbulent noise around the instant of glottal closure. The voice source within each pitch period can be viewed as consisting of two components: one corresponding to the deterministic part of excitation and the other to the random part. Thus we need to separate the excitation signal into the deterministic and random parts. In order to do this, we take the LP residual signal for two successive pitch periods using the knowledge of the glottal closure instants. A recently proposed decomposition algorithm can be used to separate the deterministic and random parts of the excitation (d'Alesandro *et al* 1995). The results are shown in figure 13. Figure 13a shows the LP residual for two successive periods, and figures 13b and c show the deterministic and random parts of the residual, respectively. The random part shows large amplitude noise signals near the glottal closure instants, and these noise signals can be attributed to the turbulent noise in the excitation. Autocorrelation functions of these three signals are shown in figure 14. The autocorrelation function of the random part (figure 14c) does not show any peak at the pitch period as in the autocorrelation function for the deterministic part (figure 14b).

6. Analysis of vocal tract system characteristics

In this section we present methods to extract the characteristics of the vocal tract system from speech signals, especially when the characteristics are changing as in dynamic sounds such as consonant-vowel combinations. The methods are based on selecting appropriate segments of speech for analysis . The selection of the segments is based on the instants of significant excitation of the vocal tract system. The analysis window is chosen starting at each of these instants. The vocal tract system is represented by formants or the resonances extracted from the short segment. Due to selection of the segments at the instants of glottal closure, the extracted formants will be consistent across successive pitch periods in voiced speech.

For voiced speech, a short (about 3 ms) segment of speech before the significant instant is considered as belonging to the open phase and the segment after the instant is considered as belonging to the closed phase. In practice there is no guarantee that there will be a distinct closed phase region. The samples after the instant of glottal closure are analysed to extract the information in that region. Likewise, the samples before the instant are analysed and the results are attributed to the open phase region. The number of samples to be considered should be less than the period between the current and the next significant instants. The minimum number of samples required depend on the order of the model and also on the method used to compute the parameters of the model. It is preferable to use as many samples as possible but larger the number the more likely that there will be a change in the vocal tract system characteristics in the analysis interval.

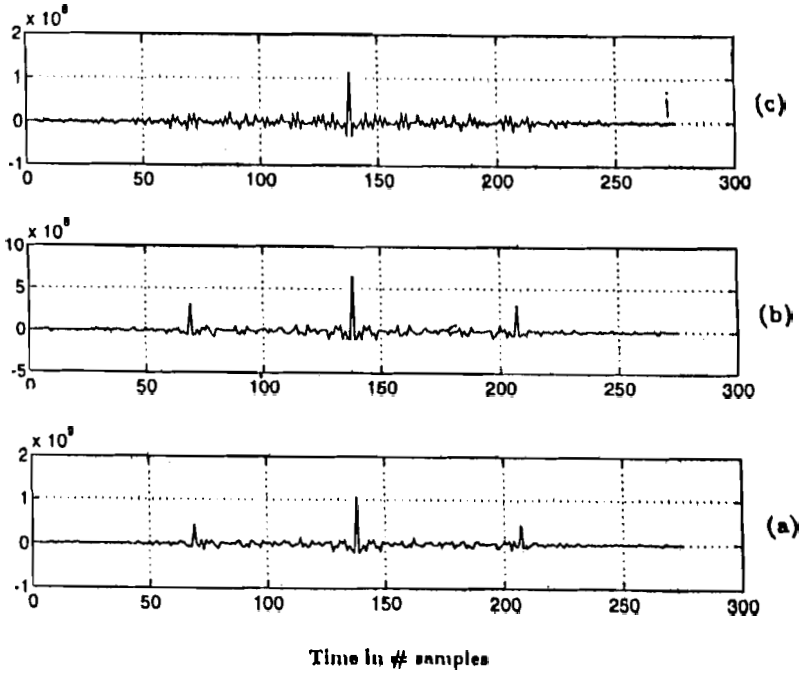

Figure 14. Autocorrelation functions of the signals in figure 13. **(a)** Autocorrelation function of the LP residual signal. (b) Autocorrelation function of the deterministic part. (c) Autocorrelation function of the random part.

A segment size of 30 samples (corresponding to 3 ms at 10 kHz sampling rate) is considered in this study. Short time spectral analysis is not useful on such segments due to poor resolution of the spectral components . High resolution model based analysis methods

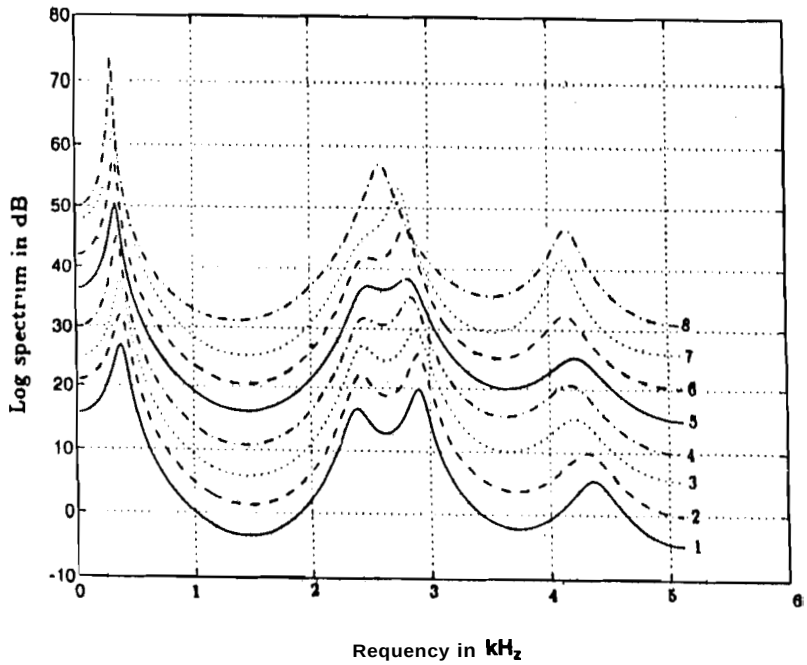

Figure 15. LP spectra (10th order) for eight successive closed phase regions in a voiced speech segment.

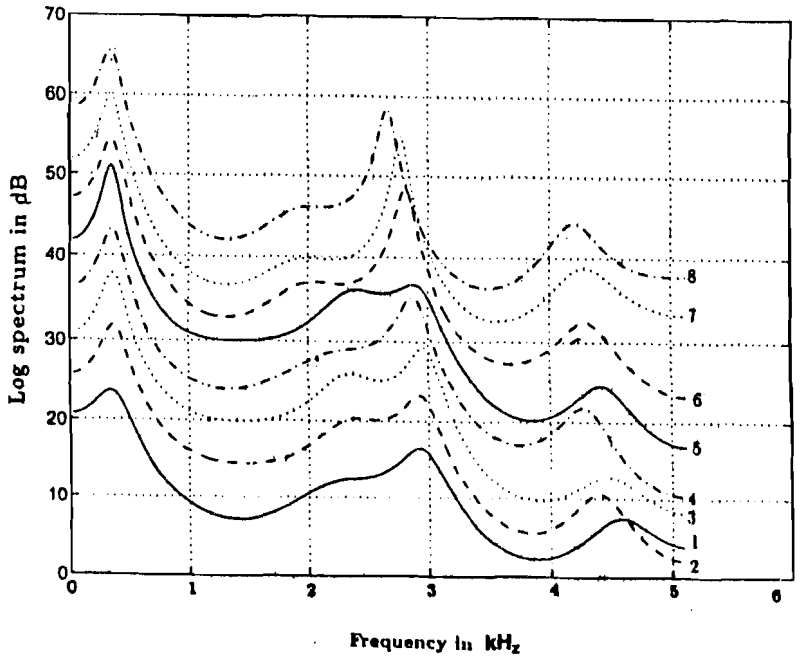

Figure 16. **LP** spectra (10th order) for eight successive open phase regions in a voiced speech segment.

can be used to extract the formant information corresponding to the vocal tract system in that region. We have used the standard covariance LP analysis to derive the LP spectrum, from which the formants can be estimated (Markel & Gray 1976).

Choosing the analysis frames synchronized with the instants of significant excitation give consistently better results than when the frames are chosen at regular spacing. Figure 15 shows the LP spectra for eight successive closed phase regions in a segment of voiced speech. The formant locations and their movements can be clearly seen from the figure. Figure 16 shows the LP spectra for the corresponding eight successive open phase regions in the segment. The formant locations and bandwidths in the open phase regions are different from the values for the corresponding closed phase regions. Moreover, the formant bandwidths are consistently higher in these open phase regions. Thus the knowledge of the instants of glottal closure for each pitch period enables us to extract the characteristics of the vocal tract system accurately.

7. Conclusions

In this paper we have discussed the importance of timing in time-frequency analysis of signals. The positioning of the analysis window in time-frequency analysis becomes critical in extracting dynamic source and system characteristics from speech signals. In order to position the analysis window suitably, it is necessary to determine the instants of significant events of production in speech signal. We have proposed a method of extracting the instants of significant excitation of the vocal tract system which corresponds to the instants of glottal closure in voiced speech. We have demonstrated that knowledge of these instants enables us to perform easily the prosodic manipulation of speech signals. It is also possible to extract the source and vocal tract system characteristics accurately, as the knowledge of the instants of significant excitation will enable us to choose the analysis segments

suitably. In particular, we have shown that the variations in the characteristics of the vocal tract system can be derived accurately by focussing the analysis on successive closed phase regions in voiced speech.

Part of this work was carried out at the Institute for Perception Research, Eindhoven Technical University, The Netherlands, during the author's visit to the Institute in 1994. The author gratefully acknowledges the contributions of Prof. Rene Collier, Mr L F Willems, Dr Raynold Veldhuis, Mr R L H Smits and Mr Remco Teunen.

References

Berkhout A J 1974 Related properties of minimum-phase and zero phase time functions. ***Geophys. Prospect.*** 22: 683–709

Childers D G, Wong C F 1994 Measuring and modelling vocal source-tract interaction. ***IEEE Trans. Biomed. Eng.*** 41: 663–671

d'Alesandro C, Yegnanarayana B, Darsinos V 1995 Decomposition of speech signals into deterministic and stochastic components. ***Proc. ICASSP*** Detroit, USA

Hams F J 1978 On the use of windows for harmonic analysis with discrete Fourier transform. ***Proc. IEEE*** 66: 51–83

Hlawatsch F, Boudreaux-Bartels G F 1992 Linear and quadratic time-frequency signal representations. ***IEEE Signal Process. Mag.***: 21-67

Markel J D, Gray A H 1976 ***Linear prediction of speech*** (New York: Springer-Verlag)

Moulines E, Laroche J 1995 Nonparameteric techniques for pitch-scale and time-scale modification of speech. ***Speech Commun.*** 16: 175–205

Oppenheim A V, Schafer R W 1989 ***Discrete-time signal processing*** (Englewood Cliffs, NJ: Prentice-Hall).

Parthasarathy S, Tufts D W 1987 Excitation-synchronous modeling of voiced speech. ***IEEE Trans. Acoust., Speech Signal Process.*** ASSP-35: 1241–1249

Rabiner L R, Atal B S, Sambur M R 1977 LPC prediction error analysis of its variation with position of the analysis frame. ***IEEE Trans. Acoust., Speech Signal Process.*** ASSP-25: 434–442

Smits R L H M, Yegnanarayana B 1994 Determination of instants of significant excitation in speech using group-delay function. Report No.886/11, IPO (Eindhoven, The Netherlands), Institute for Perception Research Manuscript (Also accepted for publication in ***IEEE Trans. Speech Audio Process.)***

Yegnanarayana B, Smits R L H M 1995 A robust method for determining instants of major excitations in voiced speech. ***Proc. ICASSP*** Detroit, USA

Yegnanarayana B, Teunen R 1994 Prosodic manipulation of speech using knowledge of instants of significant excitation. Report No.1029, Institute for Perception Research, IPO, Eindhoven, The Netherlands